

# **Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians**

---

Ronald Christensen

Department of Mathematics and Statistics  
University of New Mexico  
Albuquerque, New Mexico

Wesley Johnson

Department of Statistics  
University of California, Irvine  
Irvine, California

Adam Branscum

Departments of Public Health Oregon State  
University Corvallis, Oregon

Timothy E. Hanson

Department of Statistics  
University of South Carolina  
Columbia, South Carolina



*To Ben and Charlotte, for providing welcome, frenzied,  
and often hilarious distraction. TEH*

*To Mom, Dad, Corina, and Alex. AB*

*To Carl, for looking at me patiently, waiting for his next  
trip to the dog park, while I wrote my share of the book. I  
took great pleasure from his presence during the too  
brief time that I had it. WJ*

*To the S.I., especially Kaikoura, where I love to be with  
the people I love. RC*



---

# Contents

---

<b>Preface</b>	<b>xiii</b>
<b>1 Prologue</b>	<b>1</b>
1.1 Probability of a Defective: Binomial Data	2
1.2 Brass Alloy Zinc Content: Normal Data	3
1.3 Armadillo Hunting: Poisson Data	4
1.4 Abortion in Dairy Cattle: Survival Data	5
1.5 Ache Hunting with Age Trends	6
1.6 Lung Cancer Treatment: Log-Normal Regression	7
1.7 Survival with Random Effects: Ache Hunting	8
<b>2 Fundamental Ideas I</b>	<b>13</b>
2.1 Simple Probability Computations	14
2.2 Science, Priors, and Prediction	18
2.3 Statistical Models	22
2.4 Posterior Analysis	30
2.5 Commonly Used Distributions	33
<b>3 Integration Versus Simulation</b>	<b>37</b>
3.1 Introduction	37
3.2 WinBUGS I: Getting Started	41
3.3 Method of Composition	48
3.4 Monte Carlo Integration*	49
3.5 Posterior Computations in R	51
<b>4 Fundamental Ideas II</b>	<b>53</b>
4.1 Statistical Testing	53
4.1.1 Checking Bayesian Models	57
4.1.2 Predictive <i>P</i> -Values	59
4.1.3 Lindley-Jeffreys Paradox	60
4.2 Exchangeability	61
4.3 Likelihood Functions	63
4.4 Sufficient Statistics	66
4.5 Analysis Using Predictive Distributions	67
4.6 Flat Priors	69
4.6.1 Data Translated Likelihoods	71
4.7 Jeffreys' Priors	72
4.7.1 Multiple Parameter Jeffreys' Prior*	73
4.8 Bayes Factors*	74
4.8.1 General Parametric Testing	74
4.8.2 Nested Models	75
4.8.3 Simulating Bayes Factors	75

4.9	Other Model Selection Criteria	78
4.9.1	Bayesian Information Criterion	79
4.9.2	LPML	81
4.9.3	Deviance Information Criterion	82
4.9.4	Final Comments	83
4.10	Normal Approximations to Posteriors*	84
4.11	Bayesian Consistency and Inconsistency	88
4.12	Hierarchical Models	89
4.13	Some Final Comments on Likelihoods*	93
4.14	Identifiability and Noninformative Data	94
<b>5</b>	<b>Comparing Populations</b>	<b>97</b>
5.1	Inference for Proportions	97
5.1.1	Prior Distributions	99
5.1.1.1	Reference Priors	99
5.1.1.2	Informative Beta Priors	99
5.1.1.3	Rare Events	100
5.1.1.4	Non-Beta Priors	102
5.1.2	Effect Measures	103
5.1.3	Independent Binomials	105
5.1.4	Case-Control Sampling	107
5.2	Inference for Normal Populations	111
5.2.1	Reference Priors	111
5.2.2	Conjugate Priors	114
5.2.3	Independence Priors	115
5.2.4	Some Curious Distributional Results*	120
5.2.5	Two-Sample Normal Model	121
5.3	Inference for Rates	128
5.3.1	One-Sample Poisson Data	129
5.3.2	Informative Priors	131
5.3.3	Reference Priors	133
5.3.4	Two-Sample Poisson Data	134
5.4	Sample Size Determination*	136
<b>6</b>	<b>Simulations</b>	<b>139</b>
6.1	Generating Random Samples	139
6.2	Traditional Monte Carlo Methods	142
6.2.1	Acceptance-Rejection Sampling	142
6.2.2	Importance Sampling	143
6.3	Markov Chain Monte Carlo	145
6.3.1	Markov Chains	147
6.3.2	Gibbs Sampling	150
6.3.2.1	Proof that $p(\theta)$ is the Stationary Distribution in the Two-Block Case*	154
6.3.3	Metropolis Algorithm	154
6.3.3.1	Proof that $P(\theta)$ is the Stationary Distribution*	156
6.3.4	Slice Sampling	158
6.3.5	Checking MCMC Samples	159

## CONTENTS

ix

<b>7 Basic Concepts of Regression</b>	<b>161</b>
7.1 Introduction	161
7.2 Data Notation and Format	162
7.3 Predictive Models: An Overview	164
7.4 Modeling with Linear Structures	166
7.4.1 Continuous Predictors	166
7.4.2 Binary Predictors	166
7.4.3 Multi-Category Predictors	167
7.4.4 Predictor Selection	169
7.4.5 Several Categorical Covariates	171
7.4.6 Confounding	172
7.4.7 Effect Modification/Interaction	174
7.4.7.1 Two Categorical Predictors	175
7.4.7.2 One Continuous and One Categorical Predictor	176
7.4.7.3 Two Continuous Predictors	177
7.5 Illustration: FEV Data	178
<b>8 Binomial Regression</b>	<b>181</b>
8.1 The Sampling Model	181
8.2 Binomial Regression Analysis	186
8.2.1 Predictive Probabilities	188
8.2.2 Inference for Regression Coefficients	190
8.2.3 Inference for $LD_\alpha$	195
8.3 Model Checking	195
8.3.1 Box's Method	195
8.3.2 Link Selection	196
8.4 Prior Distributions	197
8.4.1 Simple Regression	197
8.4.2 General Regression	199
8.4.2.1 Prior Elicitation	203
8.4.2.2 Data Augmentation Priors	203
8.4.2.3 Standardized Variables	204
8.4.3 Reference Priors	207
8.4.4 Partial Prior Information	209
8.4.5 Partial Priors: Theoretical Considerations*	212
8.5 Mixed Models	213
8.5.1 Prior Elicitation	217
8.5.2 Mixed Model Likelihood	219
8.5.3 Gibbs Sampling and Centering*	219
<b>9 Linear Regression</b>	<b>223</b>
9.1 The Sampling Model	223
9.2 Reference Priors	226
9.2.1 Least Squares Estimation	227
9.2.2 Posterior Analysis	229
9.2.3 A Proper Reference Prior	230
9.3 Conjugate Priors	231
9.4 Independence Priors	233
9.4.1 Prior on $\beta$	234
9.4.2 Prior on $\tau$	236
9.4.3 Partial Prior Information	237
9.4.4 Inference and Displays	238

9.4.5	Gibbs Sampling*	239
9.4.6	WinBUGS and R Code	241
9.5	ANOVA	243
9.5.1	Independence Prior	243
9.5.1.1	Allocation and Diagnosis	248
9.5.2	Hierarchical Priors and Models	251
9.6	Model Diagnostics	252
9.7	Model Selection	257
9.8	Nonlinear Regression*	259
<b>10</b>	<b>Correlated Data</b>	<b>263</b>
10.1	Introduction	263
10.2	Mixed Models	265
10.2.1	Random Intercept Model	267
10.2.2	Random Slopes and Random Intercepts	275
10.3	Multivariate Normal Models	278
10.3.1	Parameterized Covariance Matrices	279
10.3.1.1	Analytic Formulas for CS and AR(1) Precision Matrices	283
10.4	Multivariate Normal Regression	283
10.5	Posterior Sampling and Missing Data	285
<b>11</b>	<b>Count Data</b>	<b>287</b>
11.1	Poisson Regression	287
11.1.1	Poisson Regression for Rates	289
11.2	Over-Dispersion and Mixtures of Poissons	294
11.2.1	Zero-Inflated Poisson Data	298
11.2.2	SAS Analysis of Foot-and-Mouth Disease Data	298
11.3	Longitudinal Data	300
<b>12</b>	<b>Time to Event Data</b>	<b>301</b>
12.1	Introduction	301
12.1.1	Survival and Hazard Functions	302
12.1.2	Censoring	303
12.1.3	The Likelihood	304
12.2	One-Sample Models	305
12.2.1	Distributional Models	306
12.2.2	Posterior Analysis	307
12.2.3	Log-Normal Data	307
12.2.4	Exponential Data	307
12.2.5	WinBUGS for Censored Data	308
12.2.6	Weibull Data	309
12.2.7	Prediction	311
12.2.8	Interval Censoring	312
12.3	Two-Sample Data	314
12.3.1	Two-Sample Exponential Model	314
12.3.2	Two-Sample Weibull Model	319
12.3.3	Two-Sample Log-Normal Model	320
12.4	Plotting Survival and Hazard Functions	322

## CONTENTS

xi

<b>13 Time to Event Regression</b>	<b>325</b>
13.1 Accelerated Failure Time Models	325
13.1.1 Abortion Data	333
13.1.2 Prior Elicitation for AFTs	334
13.1.2.1 Specifying the Marginal Prior for $\beta$	335
13.1.2.2 Partial Prior Information for $\beta$	338
13.1.2.3 Uncertainty About $\tau$	339
13.1.3 Case Deletion Diagnostics for AFT Models	340
13.1.3.1 Predictive Influence	342
13.1.4 Bayes Factor Model Selection	343
13.1.5 Sensitivity Analysis	343
13.1.6 Final Comments	344
13.2 Proportional Hazards Modeling	345
13.2.1 The Proportional Hazards (PH) Model	345
13.2.2 A Baseline Hazard Model	347
13.2.3 The Likelihood	347
13.2.3.1 Noninformative Data*	349
13.2.4 Priors for $\beta$	349
13.2.5 Priors for $\lambda$	351
13.2.6 Our Data Model	352
13.2.7 WinBUGS Code	353
13.2.8 Posterior Analysis for Leukemia Data	355
13.2.9 SAS Analysis of Leukemia Data	356
13.2.10 Another Example	358
13.3 Survival with Random Effects	363
<b>14 Binary Diagnostic Tests</b>	<b>365</b>
14.1 Basic Ideas	366
14.2 One Test, One Population	368
14.2.1 Gold-Standard Data	369
14.2.2 No Gold-Standard Data	371
14.3 Two Tests, Two Populations	374
14.3.1 Methods for Conditionally Independent Tests	374
14.4 Prevalence Distributions	379
<b>15 Nonparametric Models</b>	<b>385</b>
15.1 Flexible Density Shapes	386
15.1.1 Finite Mixtures	386
15.1.1.1 Identifiability Issues*	391
15.1.2 Dirichlet Process Mixtures: Infinite Mixtures	392
15.1.3 Mixtures of Polya Trees	396
15.2 Flexible Regression Functions	402
15.3 Proportional Hazards Modeling	414
<b>Appendix A: Matrices and Vectors</b>	<b>419</b>
A.1 Matrix Addition and Subtraction	420
A.2 Scalar Multiplication	420
A.3 Matrix Multiplication	420
A.4 Special Matrices	422
A.5 Linear Dependence and Rank	423
A.6 Inverse Matrices	424
A.7 A List of Useful Properties	426

A.8 Eigenvalues and Eigenvectors	426
A.9 Properties of Determinants	428
A.10 Calculus and Taylor's Theorem	428
A.11 Partitioned Matrices	428
<b>Appendix B: Probability</b>	<b>431</b>
B.1 Univariate Probability	431
B.2 Multivariate Probability	432
B.2.1 Joint Distribution of Two Vectors	434
B.2.2 Conditional Distributions	434
B.2.3 Independence	436
B.2.4 Moment Generating Functions	437
B.2.5 Change of Variables	437
B.3 Models and Conditional Independence	438
<b>Appendix C: Getting Started in R</b>	<b>443</b>
C.1 Getting R	443
C.2 Some R Basics	443
C.3 User-Contributed Packages	446
C.4 Reading Data	447
C.5 Graphing	447
C.6 Interface Between R and WinBUGS	456
C.7 Writing New R Functions	456
<b>References</b>	<b>459</b>

---

# Preface

---

Bayesian statistics embodies probability in action. What is the probability that a new chemotherapy is effective? What is the probability that a positive mammogram truly indicates breast cancer? What is the probability that a bank will become insolvent? That it will become insolvent after five years? Bayesian methods use a result known as Bayes Theorem to combine expert scientific information with data to obtain such probabilities. The famous Bayesian statistician Dennis Lindley has asserted that there are two rules in Bayesian inference: (i) always obey the laws of probability and (ii) all uncertainty is to be modeled using probability. Not surprisingly, a primary prerequisite for our exploration of Bayesian ideas and data analysis is some knowledge of probability, preferably calculus based probability.

For most of the book, probability is simply the area under a curve over a set. Such curves are probability density functions. To obtain a prediction of, say, when a new bank will become insolvent, or the mean time to insolvency for all banks, or the variability of insolvency times, Bayesian statistics requires integrating various functions against density functions. Sometimes, it is possible to evaluate integrals analytically (using calculus), and we start the book with illustrations where the calculus is tractable. However, most integrals that are necessary for Bayesian statistical inference are not tractable, so we turn to numerical approximations via simulation, in particular Markov chain Monte Carlo (MCMC) simulation. Appendices A and B review the basic concepts of matrix algebra and probability used in the book.

Of course the more statistics you know, the better for reading this book. We introduce a large number of statistical models, and the more of them that you have previously encountered, the better. Nonetheless, we have taught versions of this material to people with only a single probability course, and other versions to scientifically sophisticated students who lacked a calculus based probability course but had one or more applied statistics courses beyond a basic introduction.

The first five chapters of the book contain core material that covers basic Bayesian ideas, calculations, and inference, including modeling one and two sample data from traditional sampling models. Chapter 1 is motivational, presenting Bayesian applications from projects that we have encountered and subsequently address in the book. Chapter 2 presents the fundamentals of Bayesian philosophy and methodology including *real* prior specification, simple data models, and posterior inferences via Bayes Theorem. Chapter 3 examines the interplay between probability calculus and its approximation by computer simulation along with the implementation of simulation via the computer program WinBUGS. Chapter 4 considers deeper foundational issues. *Chapter 4 was not written thinking that everyone would read all of 4 after 3 and before 5!* It was written with the idea that people could refer to it as needed. Chapter 4 includes aspects of hypothesis testing, exchangeability, prediction, model checking and selection, “diffuse” prior specification, large sample approximations to posteriors, consistency, identifiability, and hierarchical modeling. Chapter 5 handles one and two sample analyses of binomial, normal, and Poisson data. Chapter 5 also examines relative risk estimation, case-control sampling with inferences for odds ratios, and methods for sample size determination. *Throughout the book, sections marked with an asterisk contain more sophisticated material.*

Chapter 6 introduces actual simulation methods including the theoretical basis for Markov chain Monte Carlo simulation and issues related to the practical application of MCMC.

Chapter 7 introduces general concepts of regression including linear and generalized linear modeling. Chapter 8 specifically covers binomial regression including generalized linear mixed models for correlated data. Chapter 9 presents methods for the general linear model (analysis of variance

and regression) and Chapter 10 extends those methods to handle correlated data, including longitudinal measurement data using linear mixed models and multivariate normal analysis. Chapter 11 continues the discussion of generalized linear models with coverage of Poisson regression including mixed models.

Chapter 12 introduces time to event data (survival and reliability analysis) and considers the analysis of one and two sample data that are subject to censoring. Chapter 13 continues by developing regression models for survival data, including the accelerated failure time model and Cox's proportional hazards model. The chapter concludes with a discussion of frailty models for correlated survival data.

Chapter 14 examines binary diagnostic testing. Material on continuous-response diagnostic test data appears on the book website.

Finally, Chapter 15 covers nonparametric inference, specifically density estimation and flexible regression modeling of mean functions.

The exercises form an integral part of the book. Even if readers choose not to do them, they should read them. A few of the exercises are referenced like theorems. The exercises are often located in the text to reinforce the surrounding discussion. Data, programming code, and other material can be found at the book website accessible from <http://www.math.unm.edu/~fletcher>.

We started this project with the intention of writing a compact book. We ended it with enough material for two books (Chapters 1–9 and 11 could make up one volume with the remaining chapters being a second volume). There are a number of different courses that we can envision being taught from the book. The most elementary chapters are 1, 2, 3, 5, 7, 11, and 14. If students have already had a course in regression modeling, Chapter 7 would not be necessary, although we treat it as required review reading. The most sophisticated chapters are 4, 6, 13, and 15. *Most courses would involve discussion of those parts of Chapter 4 as needed for the subsequent chapters being covered.* Chapters 4 and 6 are a must for more advanced courses. Different versions that we envision are:

- (1) M.S. Statistics Students: Chapters 1–3, 5, 8–11, plus selections from remaining chapters.
- (2) A Second M.S./Ph.D. Course: Chapters 4, 6, 12–15, plus Topics.
- (3) Ph.D. Statistics Students: Chapters 1–6, 8–9, plus selections from remaining chapters.
- (4) Biostatistics Students: Chapters 1–3, 5, plus selections from 6 and 8–14.
- (5) Epidemiology Students: Chapters 1–3, 5, plus selections from 8, 11, 12, 13, 14.
- (6) Non-Statistics Students: Chapters 1–3, 5, 7, plus selections from 8, 9, 11.

Historically, Bayesian books have been written either without analyzing real data or without discussion of the software needed to analyze data. This book is written emphasizing the use of WinBUGS and R to analyze real data. WinBUGS is introduced in Chapter 3 and R in Appendix C. It is certainly feasible to teach a Bayesian course without requiring either of these software packages. For instance, we illustrate the use of SAS in some examples. Of course it is important for statistics Ph.D. students to learn to write their own programs for analyzing unusual new data structures. Instructors can easily augment the course with assignments to write code in the language of their choice. Regardless of the programming language used, our WinBUGS code serves as a paradigm both for modeling and for the types of inferences we consider appropriate.

### Origins

In the 1980s, Statistics 145 at UC Davis was successfully taught by WJ as a purely theoretical course using the seminal book by Jim Berger. It evolved through the early 1990s, including a failed attempt to make the course accessible to non-statistics graduate students. The missing component was the lack of a convenient platform to analyze a wide variety of data. The course finally succeeded with a broad audience, including Statistics students, upon incorporating the WinBUGS software package.

All incarnations of the course have emphasized foundational issues. The immediate genesis of this project was a set of notes transcribed from that class.

Ultimately, the seeds for this book were planted in 1974 when WJ and RC became graduate students in the School of Statistics, University of Minnesota. The environment there, created in no small part by its director, Seymour Geisser, fostered examination of the foundational aspects of Statistics in virtually all areas of the program. Although there were no courses labeled “Bayesian,” Bayesian ideas suffused many of the courses. We “grew up” thinking that Bayesian methods were an integral part of inferential statistics.

The final key phase of development devolved from efforts with Ed Bedrick to specify prior information in a realistic way for generalized linear models. Lots of theoretical discussions talked about prior information, but we wanted to do something about it; in a way that really allows incorporation of input from subject matter experts.

#### *Acknowledgments*

It has been our goal to emphasize the importance of prediction, foundational issues, and ease of computation. Those goals would simply not exist except for the influence, first, of Seymour Geisser and, second, of David Spiegelhalter and the others associated with WinBUGS.

We thank Dr. Edward J. Bedrick (University of New Mexico) for his collaborations over the past 35 years, especially those involving Bayesian regression, binomial regression, and accelerated failure time models. Dr. Mark Thurmond (University of California, Davis) has been a valued collaborator providing data, subject matter expertise, and strong support of Bayesian applications in Veterinary Medicine for over 25 years. We also thank Dr. Turner Osler (formerly of the University of New Mexico), Dr. Johann Herberth (University of Kentucky), Dr. Garnett McMillan (National Center for Rehabilitative Auditory Research, Portland VA Medical Center), and Dr. Sharon Hietala (University of California, Davis) for providing data, and Dr. Osler, Dr. McMillan, and Dr. David Mannino (University of Kentucky) for their efforts in quantifying their expert opinions and in explicating the background, procedures, and goals of their studies. We also thank Dr. Ian Gardner (University of California, Davis) for his long-standing collaboration and support of Bayesian applications in Veterinary Epidemiology. Dr. David Fardo provided valuable feedback on Appendix C. Fletcher Christensen played Strunk to Ron’s White on any elements of style and contributed substantially to the treatment of prior information for rare events.

One of us is embarrassed by the number of times he has referenced himself. He claims that it is more out of sloth than egotism, not that he has any lack of either.

Any errors in the book are Carl’s fault. He ate much of Wes’s homework!

Ronald Christensen

Orin Wesley Johnson

Adam Branscum

Tim Hanson



---

## Chapter 1

---

# Prologue

---

In this book our intention is to emphasize:

- I. That a primary role of statistics in society is to provide appropriate tools for addressing *scientific* questions.
- II. That appropriate statistical analysis of data involves a collaborative effort between subject matter scientists and statisticians.
- III. That it is both appropriate and necessary to incorporate the scientist's expertise into making decisions related to the data.
- IV. That foundational issues matter in statistics.
- V. That prediction is of fundamental importance. Science and technology are about making accurate predictions. They are about being able to say that in certain circumstances, certain things will happen. Objectivity and understanding are only of permanent scientific value when they lead to testably accurate predictions.

To achieve these aims, we have adopted a Bayesian approach. Bayesian statistical analysis is based on the premise that all uncertainty should be modeled using probabilities and that statistical inferences should be logical conclusions based on the laws of probability.

The field of Statistics has long embraced the concept of probability models for data. Such models typically involve parameters that are presumed to be related to characteristics of the sampled populations. These parameters can range from few in number with simple interpretations to an uncountable number. Parameters can never be known with absolute certainty unless we sample the entire population. Moreover, parameters may not have physical interpretations since, inevitably, models rarely are precisely true. Models are, we hope, useful approximations to some truth that can provide good predictions. Nonetheless, statistical inquiry has focused more on the estimation of parameters than on prediction.

Given a statistical model for the data, the Bayesian approach mandates an additional probability model for all unknown parameters in the data model. Our approach is to model this uncertainty about the parameters using scientific expert information. This information is called "prior" information, or information that has been collected *a priori*. Expert information must be obtained independently of the data being analyzed. One way to guarantee that scientific input about model parameters is independent of the data is to acquire that information before the data have been collected. However, despite the *a priori* terminology, such information is often not literally obtained prior to the collection of data. Our experience is that it is generally possible to obtain independent information from sources such as existing literature or colleagues of the scientists who collected the current data. Throughout the book, we use the word "prior" partly for simplicity of exposition and partly for historical reasons, but it is understood that prior information is simply information obtained independently of the current data.

In the remainder of this chapter we provide a taste of the types of data and problems that Bayesian methods can address. In the remainder of the book we present information needed to develop the results in this chapter. The chapter starts with simple probability models for binomial, normal, and Poisson data. It works up to more elaborate models including regression models and

models that incorporate random effects. In some of these examples we have prior information that we incorporate into the model. In others we use *reference priors* that provide a basis for discussion and comparison. Typically, the reference priors incorporate little prior information and to some extent approximate non-Bayesian approaches.

### 1.1 Probability of a Defective: Binomial Data

The Par-Aide Corporation in Lino Lakes, Minnesota, makes ball washers for golf courses. St. Paul Brass and Aluminum Foundry makes a part called a “push rod eye,” an integral component of the golf ball washer. Out of 2,430 push rod eyes poured over seven days in May, 2003, only 2,211 actually shipped. (It was really only two days, but “Seven ...” is such a good book.) It is of interest to estimate the probability of pouring a defective part. The Vice President of Operations thinks that for this particular part a plausible range for the proportion of scrap is 5% to 15%.

We assume that the production process determines a proportion of defective parts that we call  $\theta$ . We treat the available data as a random sample from the production process, so that the probability of any part being scrapped is  $\theta$  and further we assume that all 2,430 parts being examined are independent. With these assumptions, the number of defective parts  $y$  has a binomial distribution. Write

$$y|\theta \sim \text{Bin}(2430, \theta).$$

This is the *sampling distribution*.

The proportion defective,  $\theta$ , is unknown. We use probability to reflect the Vice President’s knowledge about it. We interpreted the Vice President as specifying  $\Pr(\theta \leq 0.05) = 0.025$  and  $\Pr(\theta \geq 0.15) = 0.025$ . Although these statements do not completely determine a probability distribution for  $\theta$ , we restricted the probability distribution on  $\theta$  to fall into a convenient but relatively flexible class, and within that class the two statements determine a complete probability distribution for  $\theta$ . The distributions are called beta distributions, so we write

$$\theta \sim \text{Beta}(12.05, 116.06),$$

where the numbers 12.05 and 116.06 were chosen so that the distribution agrees with the Vice President’s two statements. This is the *prior distribution*.

A Bayesian analysis uses a result called *Bayes Theorem* to combine the data with the prior distribution to update the Vice President’s probability distribution about  $\theta$ . This new probability distribution, called the *posterior distribution*, describes knowledge about  $\theta$  and is the fundamental tool in Bayesian statistical analysis. Typically, we use computer simulations to approximate the posterior distribution. Occasionally, we can find it mathematically.

With  $y = 219$ , computer simulations give the estimated median of the posterior distribution as  $\tilde{\theta} = 0.09$ . There is a 95% probability that  $\theta$  falls within  $(0.08, 0.10)$ , with only a 2.5% chance that  $\theta$  is either smaller than 0.08 or larger than 0.10. This is a 95% *equal-tailed probability interval (PI)*. (*Unless otherwise mentioned, all our probability intervals will be equal-tailed.*)

In this model it is possible to obtain the posterior distribution mathematically. As shown in Section 2.3, if

$$y|\theta \sim \text{Bin}(n, \theta)$$

and

$$\theta \sim \text{Beta}(a, b),$$

then

$$\theta|y \sim \text{Beta}(y + a, n - y + b).$$

For the data of our example, the posterior distribution is  $\text{Beta}(219 + 12.05, 2430 - 219 + 116.06)$  with median 0.0902 and probability interval  $(0.0795, 0.1017)$ . These exact calculations agree with the approximations obtained from computer simulation to two decimal places.

In analyzing these data, we made a number of assumptions, particularly with the prior. It is wise to do a sensitivity analysis to see how much the conclusions depend on the assumptions.

## 1.2 Brass Alloy Zinc Content: Normal Data

A corrosion resistant brass alloy, widely used in plumbing fixtures, is composed mainly of copper, tin, lead, zinc, nickel, and iron in decreasing amounts. The addition of zinc to the alloy produces a jump in metal strength when added within a tolerance range between 4% and 6%. Zinc has a lower melting point than copper, so a certain amount of the added zinc is dissipated by the end of the heating process. It is common practice to measure the actual zinc content through spectrometry readings before pouring the metal alloy into molds. If the amount of zinc is less than, typically, 4.4%, zinc is added to the alloy to correct the percentage.

Twelve alloy samples were tested using spectrometry by the St. Paul Brass and Aluminum Foundry in St. Paul, Minnesota, in June of 2003. Let  $y_i$  be the zinc percentage of sample  $i$ ,  $i = 1, \dots, 12$ . We assume that the observations are independent and follow a bell-shaped curve. Specifically, we assume they follow a normal distribution with some mean  $\mu$  and variance  $\sigma^2$ . Write

$$y_1, \dots, y_{12} | \mu, \sigma^2 \stackrel{iid}{\sim} N(\mu, \sigma^2).$$

The *iid* in the relation indicates that the observations are *independent and identically distributed*. The  $n = 12$  sample zinc percentages were 4.20, 4.36, 4.11, 3.96, 5.63, 4.50, 5.64, 4.38, 4.45, 3.67, 5.26, and 4.66. A histogram and normal quantile plot (see Christensen, 1996, Sec. 2.4) show no serious deviations from the assumption of a normal distribution.

To perform a Bayesian analysis, we need to identify a prior distribution that gives information about the unknown parameters  $\mu$  and  $\sigma^2$ . The Vice President of Operations, Kay Stinson, estimates with 95% certainty that the mean percentage before pouring should be between 4.5% and 5%, and centered at 4.75%. We interpret these ideas as requiring that the prior distribution have expected value  $E(\mu) = 4.75$  and  $\Pr(4.5 < \mu < 5) = 0.95$ . Two such statements do not determine a prior distribution, but as in the previous section, we restrict the class of possible prior distributions so that they will. For normal data, the convenient prior distribution on  $\mu$  is also a normal distribution with some mean and variance. To satisfy the Vice President's specifications, we need

$$\mu \sim N(4.75, 0.0163).$$

We have no good information on the variance  $\sigma^2$ , so we specify a “reference prior” on  $\sigma^2$  that is independent of  $\mu$ . By a *reference prior* we mean any prior that is chosen to provide a common base for people to evaluate data, rather than one that models specific prior information. The reference prior we use is a gamma distribution on  $1/\sigma^2$  with very small parameters. In particular, we used

$$\sigma^{-2} \sim \text{Gamma}(0.001, 0.001).$$

For the parameter  $\mu$ , computer simulations gave the posterior median as  $\tilde{\mu} = 4.69\%$ , that is,  $\Pr(\mu \leq 4.69 | y_1, \dots, y_{12}) = 0.5$ . For  $\sigma$  the median is  $\tilde{\sigma} = 0.64\%$ . Similarly, there is a probability of 0.95 that  $\mu$  is between 4.49% and 4.90% with probabilities of 0.025 of being smaller than 4.49 or larger than 4.90. In particular,  $\Pr(4.49 < \mu < 4.90 | y_1, \dots, y_{12}) = 0.95$ . For  $\sigma$  the corresponding 95% probability interval is (0.44%, 1.03%).

It is of additional interest to find the probability that zinc needs to be added to the metal before casting. Let  $y_{13}$  be the zinc score of a future batch. We want to know if  $y_{13} \leq 4.4$ . To evaluate this, we assume

$$y_{13} | \mu, \sigma \sim N(\mu, \sigma^2)$$

independent of  $y_1, \dots, y_{12}$  (given  $\mu$  and  $\sigma$ ). Our new parameter of interest is  $\Pr(y_{13} \leq 4.4 | \mu, \sigma)$  and our estimate of it is the predictive probability  $\Pr(y_{13} \leq 4.4 | y_1, \dots, y_{12})$ . From computer simulations the probability of needing to add zinc is 0.33 with a 95% probability interval for  $\Pr(y_{13} \leq 4.4 | \mu, \sigma)$  of (0.20, 0.45).

In analyzing these or any other data, we make a number of assumptions for both the sampling distribution and the prior. It is always wise to do a sensitivity analysis to see how much the conclusions depend on the assumptions.

### 1.3 Armadillo Hunting: Poisson Data

The Ache tribe of Paraguay are part-time hunter-gatherers and have been in contact with Paraguayan society only since the mid-1970s. McMillan (2001) collected data on many aspects of Ache life including hunting, nutrition status, demographic features, child care, and health. Part of Ache life is spent away from the village on extended forest treks. Each trek lasts from three days to several weeks. Between one and six families participate. While on trek, men typically search for game animals by themselves but come together to capture peccaries, monkeys, and paca in groups. While trekking, the Ache subsist exclusively on foods that they collect on a given day. Meats are agouti paca, peccary, deer, coati, capuchin monkeys, a variety of reptiles, toucans and other birds, and insects. Gathered foods include palm hearts, palm starch, oranges and other fruit, and honey. However, armadillos comprise the vast majority of food calories consumed by the Ache and it is of interest to quantify the typical number of armadillos killed in a day.

Our data involve observations on  $n = 38$  Ache men. Let  $y_i$  be the number of armadillos killed by the  $i$ th man on a given day. The Poisson distribution provides a natural model for the number of events occurring haphazardly over a given amount of time, so we assume

$$y_1, \dots, y_n | \theta \stackrel{iid}{\sim} \text{Pois}(\theta).$$

We refer to  $\theta$  as the *kill rate*. (Something that should also be analyzed for Vin Diesel movies.) The broader Ache data include many hunting days for each man. For this analysis one day was randomly selected for each man.

A gamma prior distribution conveniently provides a model for prior information on the mean daily number of kills  $\theta$ . Dr. Garnett McMillan, an expert on Ache hunting practices, believes that Ache men typically kill an armadillo every other day and thus provides a “best guess” for  $\theta$  of 0.5 armadillos, which we take to be the median of the prior distribution. Dr. McMillan is 95% sure that the mean daily number of kills is no greater than 2 armadillos. With

$$\theta \sim \text{Gamma}(a, b),$$

we solve the (rather complicated) simultaneous equations

$$\Pr(\theta \leq 0.5 | a, b) = 0.50, \quad \Pr(\theta \leq 2 | a, b) = 0.95$$

for  $a$  and  $b$  yielding  $a = 1.11$  and  $b = 1.61$ . Thus our prior is

$$\theta \sim \text{Gamma}(1.11, 1.61).$$

In this simple model the posterior for  $\theta$  may be found mathematically. It turns out that

$$\theta | y_1, \dots, y_n \sim \text{Gamma}\left(\sum_{i=1}^n y_i + a, n + b\right).$$

In the case of the Ache data,  $n = 38$ ,  $\sum_{i=1}^{38} y_i = 10$ , so with  $a = 1.11$  and  $b = 1.61$  we obtain

$$\theta | y_1, \dots, y_{38} \sim \text{Gamma}(11.11, 39.61).$$

Whereas the prior median daily kill rate was  $0.497 \doteq 0.5$  armadillos per day with a 95% probability interval of  $(0.024, 2.433)$ , the posterior median daily kill rate is 0.272 armadillos per day with a 95% probability interval of  $(0.140, 0.468)$ . In other words, after incorporating the data, with probability 0.95, the mean number of armadillos killed per day is between 0.14 and 0.47 armadillos, or one armadillo killed per 2 to 7 days.

In this example the posterior focuses more tightly on smaller values of  $\theta$  than the prior, see Figure 1.1. The data indicate that the mean number of kills is less than the expert’s best guess. In

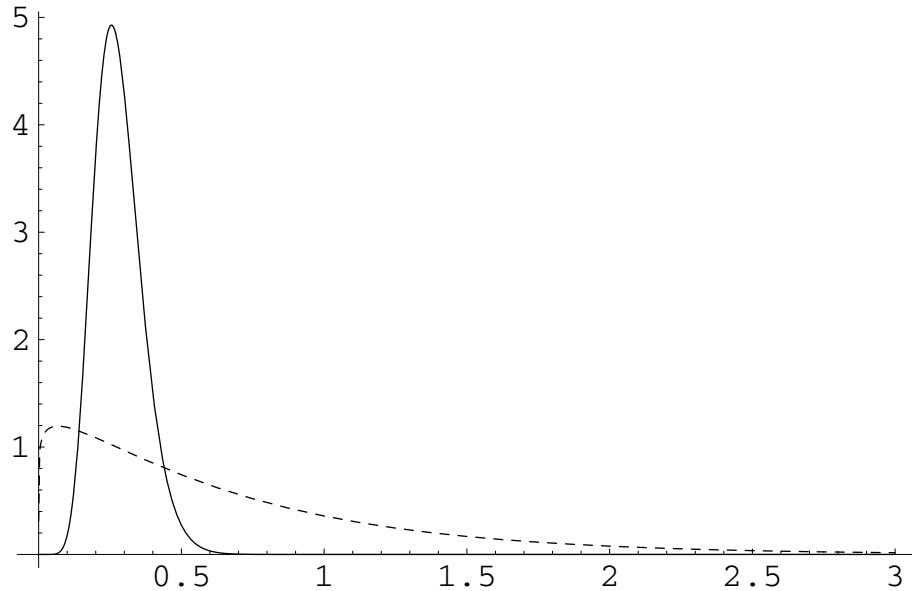


Figure 1.1: Prior (dashed) and posterior (solid) distributions for kill rate  $\theta$ .

fact, the posterior 95% probability interval for  $\theta$  does not contain 0.5, so we could reasonably reject the value 0.5 as implausible. Does this suggest that the expert's opinion is suspect? Not at all! The expert's prior encompasses a wide range of plausible values for  $\theta$  that are roughly centered at 0.5 and the posterior 95% probability interval falls well within the middle 95% of the prior.

If the posterior probability interval had fallen far outside the prior interval we would have evidence of a discrepancy; the data and the expert would indicate two very different scenarios regarding the daily kill rate of armadillos. This could happen, for example, if the expert's opinion was based on the historical abundance of armadillos and the population declined shortly before the sample was collected. Such a divergence suggests rethinking of armadillo abundance or Ache hunting habits by the expert. A strength of the Bayesian paradigm is that it provides a natural forum for the comparison and synthesis of current data with scientific opinion derived from theory and historical information.

As always, a sensitivity analysis would be wise.

#### 1.4 Abortion in Dairy Cattle: Survival Data

Our next example is an illustration of *time to event data*. These are data about how long it takes for something to happen. How long until your refrigerator breaks down? How long until a drug kicks in? How long does a person with a fatal disease survive? Such data often involve a complicating factor called *censoring*. Some people buy a refrigerator because theirs broke. Those people know how long their refrigerator survived. But many people replace their refrigerator before it breaks, so the data on how long it would last is censored. All they know is that it was still running when it got replaced. The following simple example does not involve censoring but censoring is discussed briefly in Section 1.6 and more extensively in Chapter 12.

Bedrick, Christensen, and Johnson (2000) consider data on 45 cows that naturally aborted their fetuses prematurely. It is of interest to dairy managers to determine whether cows infected with *Neospora caninum* typically abort later than uninfected cows; 19 of the 45 cows were infected. The times to abortion in the uninfected group are 60, 74, 37, 45, 75, 40, 50, 50, 146, 70, 50, 84, 60, 149, 50, 90, 259, 40, 90, 101, 70, 90, 254, 130, 80, and 40 days. For the infected group the times are 50, 130, 100, 130, 50, 140, 129, 76, 138, 69, 70, 144, 70, 130, 70, 150, 251, 110, and 120 days.

Let  $y_{1j}$  be the time to abortion of the  $j$ th cow infected with *Neospora caninum* and let  $y_{2j}$  be the time to abortion of the  $j$ th cow in the uninfected group. We assume that the logs of the abortion times in each group are normal with means  $\mu_1$ ,  $\mu_2$  and variances  $\sigma_1^2$ ,  $\sigma_2^2$ . Referring to the abortion times as having *log-normal distributions*, write

$$y_{11}, \dots, y_{1,19} | \mu_1, \sigma_1^2 \stackrel{iid}{\sim} LN(\mu_1, \sigma_1^2)$$

and

$$y_{21}, \dots, y_{2,26} | \mu_2, \sigma_2^2 \stackrel{iid}{\sim} LN(\mu_2, \sigma_2^2).$$

We use independent reference priors

$$\mu_i \sim N(0, 1000) \quad \text{and} \quad 1/\sigma_i^2 \sim \text{Gamma}(0.001, 0.001), \quad i = 1, 2.$$

Given the model parameters, the median times to abortion in the two groups are  $\exp(\mu_1)$  and  $\exp(\mu_2)$  so the difference in medians is  $\Delta = \exp(\mu_1) - \exp(\mu_2)$ . We are interested in the posterior distribution  $\Delta | y_{11}, \dots, y_{1,19}, y_{21}, \dots, y_{2,26}$  and whether the difference is reasonably close to zero.

Computer simulations were used to estimate the posterior median of  $\Delta$ , which is 27.8 days, and the equal-tailed 95% probability interval for  $\Delta$ , which is (1.7, 55.2) days. With 95% probability, infected cows have a median time to abortion between 1.7 and 55.2 days longer than those cows not infected with *Neospora caninum*. If we use alternative reference priors that are diffuse but finite uniform distributions on the means and precisions, these numbers change only slightly; the posterior median is 28.0 days and the 95% probability interval is 1.8 to 54.9 days.

### 1.5 Ache Hunting with Age Trends

We now return to Ache hunting with more data and a more sophisticated model that incorporates random hunter effects into the analysis along with a tendency for hunting success to change with age. Data were collected on the daily number of armadillos killed by 38 adult males of an Ache tribe over several forest treks. There are 1,302 total observations  $y_{ij}$  where  $i = 1, \dots, 38$  indexes an Ache male and  $j = 1, \dots, n_i$  denotes a day spent hunting. A plot of the average number of kills per man by age shows a generally increasing, then decreasing trend. It is of interest to model and quantify how a man's age affects daily kill success. We assume the number of armadillos killed is distributed Poisson and take the log-rate to be a quadratic function of a man's age in years  $a_i$  plus a subject-specific, normally distributed random effect  $\delta_i$ . The random effect for each man accounts for the correlation of an individual's daily kills over the several hunting trips in which the data were collected. One might also think of  $\delta_i$  as the innate ability of hunter  $i$ .

The sampling model is specified

$$\begin{aligned} y_{ij} | \lambda_i &\stackrel{ind}{\sim} \text{Pois}(\lambda_i), \quad i = 1, \dots, 38; j = 1, \dots, n_i, \\ \log(\lambda_i) &= \beta_1 + \beta_2(a_i - \bar{a}) + \beta_3(a_i - \bar{a})^2 + \delta_i \\ \delta_i | \tau &\stackrel{iid}{\sim} N(0, \tau^{-1}). \end{aligned}$$

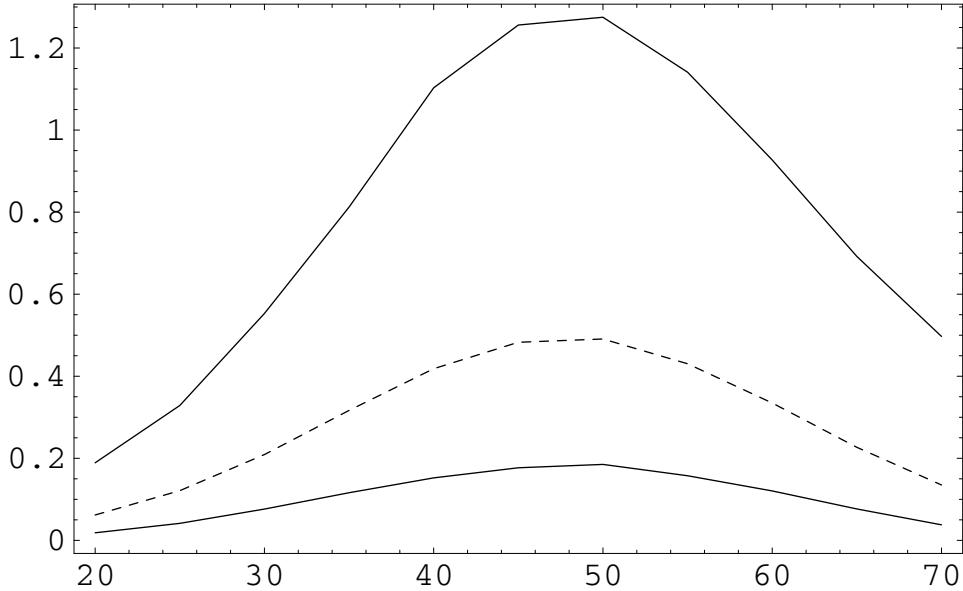
Here  $\lambda_i$  is the mean daily kill rate for individual  $i$ ,  $\bar{a}$  is the average age of the 38 hunters, while the variance of the normal distribution, usually denoted  $\sigma^2$ , has been reparameterized as the *precision*

$$\tau \equiv \frac{1}{\sigma^2},$$

which turns out to be much more convenient for Bayesian analysis of normal distributions. The ages  $a_i$  are fixed, known constants. The parameters  $\beta_1, \beta_2, \beta_3$ , and  $\tau$  are given independent reference prior distributions  $\beta_i \sim N(0, 1000)$ ,  $i = 1, 2, 3$  and  $\tau \sim \text{Gamma}(0.001, 0.001)$ .

Table 1.1: *Armadillo kills: posterior medians and probability intervals.*

Parameter	Median	95% PI
$\beta_1$	-0.7147	(-1.007, -0.433)
$\beta_2$	0.01368	(-0.002525, 0.03085)
$\beta_3$	-0.002683	(-0.004007, -0.001459)
$\sigma$	0.4252	(0.2731, 0.658)

Figure 1.2: *Estimated mean daily kill by age with 95% PI.*

We obtain posterior information by simulation. Summary information is presented in Table 1.1. The quadratic coefficient  $\beta_3$  is estimated to be  $-0.0027$  and from the probability interval it is clearly nonzero. If we exponentiate  $\log(\lambda)$ , we obtain an estimate of the mean daily kill rate as a function of age  $a$ ,

$$\hat{\lambda}(a) = \exp\{-0.7147 + 0.01368(a - \bar{a}) - 0.002683(a - \bar{a})^2\},$$

although this estimate is not necessarily the posterior median.

Posterior medians for the mean daily kill  $\lambda(a)$  and probability intervals were computed for ages 20 to 70 in steps of five years to obtain Figure 1.2. The range of ages in the data is 20 to 66 years. We see that the average kill-rate increases with age up until about 50, perhaps reflecting that hunting experience increases the chance of killing an armadillo, but then declines as the hunter enters his “golden years.” When the model was refit using independent, infinite, and improper priors on all model parameters the resulting posterior inferences were almost identical to those presented above.

## 1.6 Lung Cancer Treatment: Log-Normal Regression

The drugs cisplatin and etoposide can increase the lifetimes of those with limited-stage small cell lung cancer. It was of interest to determine which sequencing of the drugs works better. Treatment 0 was the administration of cisplatin followed by etoposide. Those receiving treatment 1 were given etoposide followed by cisplatin. The 121 patients studied were randomly assigned to the two treatment groups: 62 patients received treatment 0 and 59 patients received treatment 1. The data are the time in days  $y$  that a patient was known to be alive from the start of the treatment regimen, along

Table 1.2: *Cancer treatments: posterior medians and probability intervals.*

Parameter	Median	95% PI
$\beta_1$	7.7	(6.7, 8.7)
$\beta_2$	-0.0184	(-0.0344, -0.0028)
$\beta_3$	-0.4024	(-0.6914, -0.1210)
$\tau$	1.72	(1.25, 2.30)

with the explanatory variables treatment group ( $g$ ) and patient age ( $A$ ) at entry into the study. Define  $T$  to be the time from the start of a treatment regimen until death due to small cell lung cancer. The *survival time*  $T$  may or may not be the recorded time  $y$ . For some individuals the survival time  $T$  is *right censored*. That is, the survival time is known only to be greater than the time recorded:  $T > y$ . This can happen for several reasons. For example, if a patient is alive at the end of the study we know only that survival time is longer than the time the individual spent in the study. Each individual has a noncensoring indicator  $\delta$  that is 0 for a right censored observation ( $T > y$ ) and 1 for an uncensored observation ( $T = y$ ). A portion of the data is reproduced here:

$i$	$A_i$	$y_i$	$\delta_i$	$g_i$
1	56	730	1	0
2	70	1980	0	0
:	:	:	:	:
121	63	254	1	1

With  $T_i$  the survival time for patient  $i$ , consider the log-normal regression sampling model

$$\begin{aligned} \log T_i &= \beta_1 + \beta_2 A_i + \beta_3 g_i + \varepsilon_i, \quad i = 1, \dots, 121, \\ \varepsilon_i | \tau &\stackrel{iid}{\sim} N(0, 1/\tau). \end{aligned}$$

With  $\perp\!\!\!\perp$  denoting independence, we use reference priors

$$\beta_1, \beta_2, \beta_3 \stackrel{iid}{\sim} N(0, 1000) \quad \perp\!\!\!\perp \quad \tau \sim \text{Gamma}(0.001, 0.001).$$

We have placed approximately flat (uniform) distributions on  $\beta_1$ ,  $\beta_2$ , and  $\beta_3$  by specifying a large variance in the independent normal priors for these parameters.

Figure 1.3 illustrates survival probabilities. Posterior estimates, based on simulations, are given in Table 1.2. Treatment group 1 survives an estimated  $\exp(-0.4024) = 0.67$  times as long as group 0. In other words, among patients with the same age at entry, we estimate that the life expectancy for people on treatment 1 is only two thirds of that for people on treatment 0. The 95% PI for  $\beta_3$  contains only negative values so we are confident that people in treatment group 1 do not survive as long as people in group 0. In either treatment group, adding ten years to one's entry age decreases median survival time by a multiplicative factor of about  $\exp(-0.0184 \times 10) = 0.83$ .

## 1.7 Survival with Random Effects: Ache Hunting

We continue our examination of Ache hunting skill by looking at the time it takes to find an armadillo. Much current anthropologic theory assigns great importance to skill in hunting, which varies from person to person due to experience, age, and cognitive ability. Furthermore, an Ache male's status within the group is highly determined by his skill in hunting.

A search time  $T$  is defined to be the time in minutes from when an Ache hunter starts walking to when an armadillo burrow is found. The search is right censored if it stops before finding an armadillo burrow. This may occur because other animals are encountered, the hunter takes a break, or darkness occurs and the hunter settles in camp for the night. For right censored observations, we

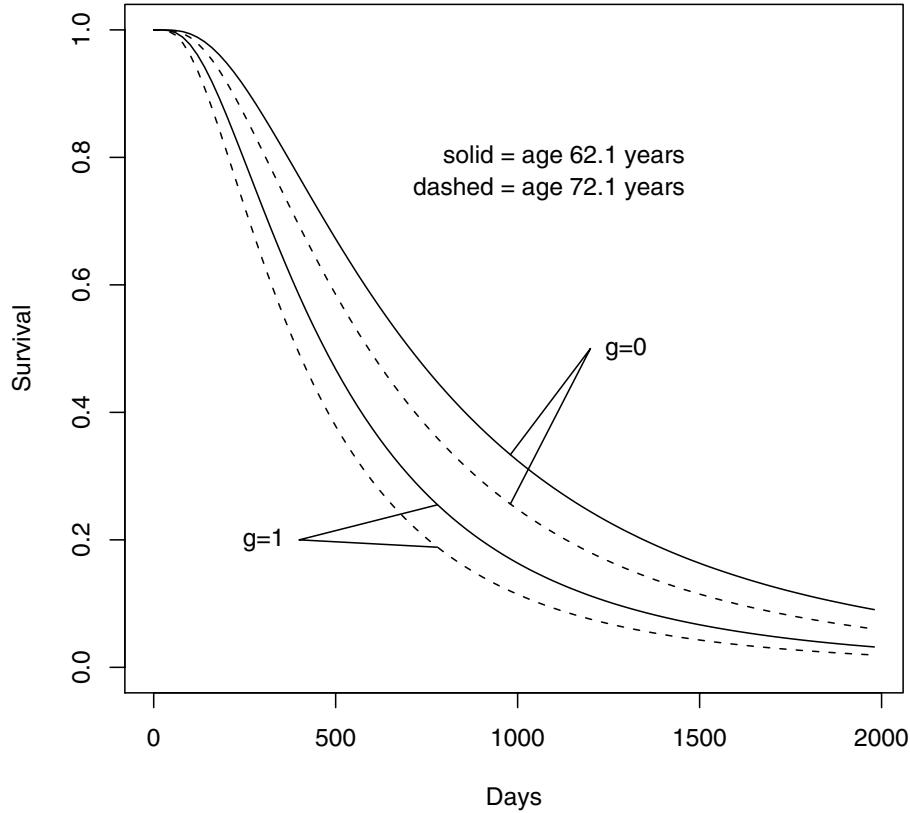


Figure 1.3: *Lung cancer survival as functions of age, treatment and time.*

only know that the true search time would have been longer than the recorded time had the hunter not been distracted from his task.

Let  $i = 1, \dots, 14$  denote a specific Ache male, while  $j = 1, \dots, n_i$  indexes a particular attempt at finding an armadillo burrow. We consider a model in which the log of the search time  $T$  is normally distributed, i.e.,

$$\log(T_{ij}) = \mu + \delta_i + \varepsilon_{ij}, \quad (1)$$

where  $\mu$  is an overall search effect,  $\delta_i$  is a random effect for the skill of the  $i$ th hunter and  $\varepsilon_{ij}$  is a random effect for a particular search. The model further specifies

$$\varepsilon_{ij} | \sigma \stackrel{iid}{\sim} N(0, \sigma^2) \quad \perp \quad \delta_i | \sigma_\delta \stackrel{iid}{\sim} N(0, \sigma_\delta^2). \quad (2)$$

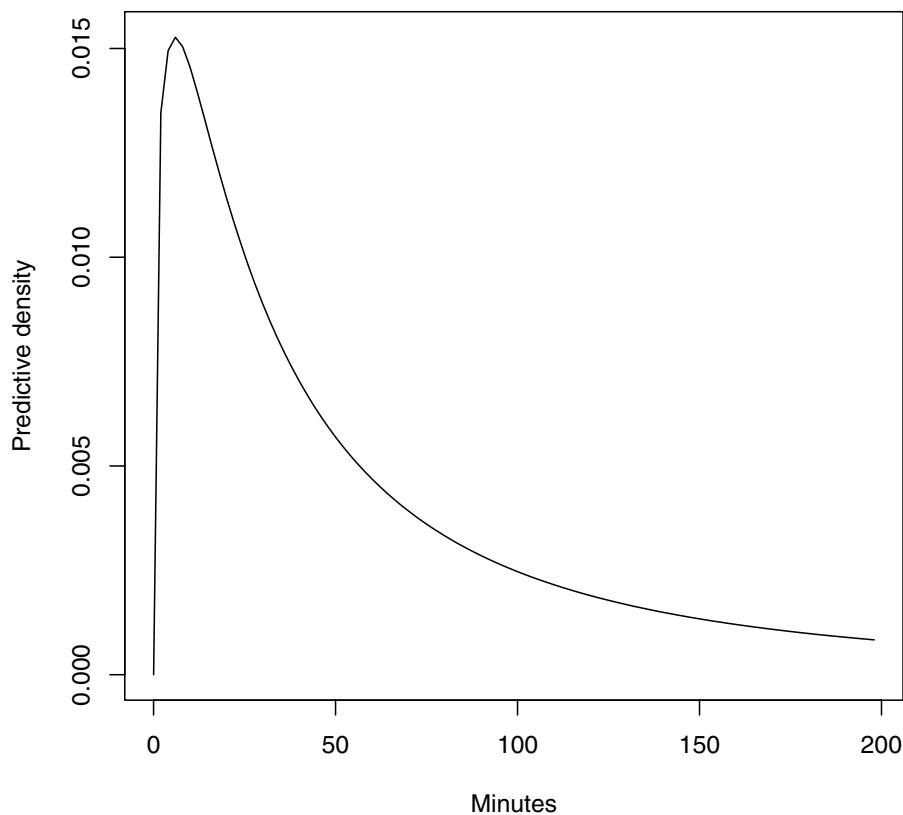
The log survival times follow a normal distribution, that is, the survival times follow a log-normal distribution. Furthermore, we have assumed that the *acceleration factors*  $e^{\delta_i}$  follow a log-normal distribution as well. (More on this later.)

We elicited priors for this model from our Ache expert, Dr. McMillan, assuming that model parameters are *a priori* independent. As discussed in the following subsection, the prior turns out to be  $\mu \sim N(4.5, 0.1)$ ,  $\sigma \sim \text{Gamma}(2.29, 2.92)$ , and  $\sigma_\delta \sim \text{Gamma}(8.1, 24.5)$ .

Define  $T_f$  to be the search time for a randomly selected Ache male different from the ones included in the analysis. In Table 1.3 we present posterior results for model parameters and  $T_f$ .

Table 1.3: *Prior and posterior medians and 95% probability intervals.*

Parameter	Prior		Posterior	
	Median	95% PI	Median	95% PI
$\mu$	4.5	(3.9, 5.1)	3.9	(3.6, 4.2)
$\sigma$	0.67	(0.11, 2.08)	1.28	(1.11, 1.50)
$\sigma_\delta$	0.32	(0.14, 0.60)	0.36	(0.18, 0.60)
$T_f$	90	(10, 796)	48	(3, 697)

Figure 1.4: *Predictive density for finding an armadillo.*

From the predictive distribution, the median time to find an armadillo burrow is estimated to be 48 minutes with a plausible range of 3 to 697 minutes (3 minutes to about half a day). Figure 1.4 shows much of the predictive density as a function of search time. Note that the modal predictive search time is much smaller than the median of 48 minutes and that positive densities occur past 700 minutes. In a more extensive analysis, neither the hunter's age nor whether the group is within a day's walk of the village are important predictors of search-time.

This simple parametric survival model fits the data quite well. We examine survival models in Chapters 12 and 13, and in Chapter 15 include very general (nonparametric) models based on Dirichlet process mixtures and Polya trees.

### *Finding the Prior*

We have three parameters on which to find a prior distribution:  $\mu$ ,  $\sigma$ , and  $\sigma_\delta$ . To simplify matters, we assume that the parameters can be regarded independently. To further simplify matters, we assume that an adequate prior can be found by restricting the distribution of each parameter to a parametric family. For  $\mu$ , the family consists of normal distributions and for the two standard deviations the families consist of gamma distributions. Each of these are two parameter families, so if we can make two probability statements about each parameter, we can find a corresponding distribution in the family that gives those probabilities. Finding the corresponding probability distributions is somewhat computationally intensive. We focus on the process of eliciting the two probability statements. While this is a fairly simplistic approach to eliciting a prior distribution, even this can be difficult to execute. The key point with any processes of eliciting a prior distribution is to reconfirm with the expert that the prior in fact gives a reasonable representation of their prior information. Thus, when we come up with a gamma distribution for a standard deviation, it is incumbent upon us to explain to the expert, in terms they can understand, the implications of that particular prior.

Statisticians often develop rather abstruse parameters for their statistical models. Experts are generally much more familiar with potential observations than with statistical parameters. To elicit useful information about parameters, we take the position that parameters must be quantities that experts can think about easily. In the current example, rather than trying to elicit information directly about standard deviations, we elicit information about percentiles of random variables, and induce the priors from those percentiles.

The sampling distribution for the data was specified in (1) and (2). It follows from these that

$$\log(T)|\mu, \sigma, \sigma_\delta \sim N(\mu, \sigma_\delta^2 + \sigma^2).$$

The parameter  $\mu$  is both the mean and the median of the distribution. If we look at  $T = e^{\log(T)}$ , the median is  $e^\mu \equiv M$ , although this is not the mean. We will focus on eliciting information about  $M$ .

To learn about  $\sigma$ , we elicit information about an “average” hunter, one with the value  $\delta_i = 0$ . From equation (1) it follows that

$$\log(T)|\mu, \sigma, \delta_i = 0 \sim N(\mu, \sigma^2).$$

We determined that something our expert could think about was the 75th percentile of the distribution of  $T$ , given that he already knew the median. The 75th percentile is the number below which exactly 3/4s of the observations fall. For normal data like  $\log(T)$ , the 75th percentile is known to be  $\mu + z_{0.75}\sigma$  where  $z_{0.75} = 0.6745$  is the 75th percentile of a  $N(0, 1)$  distribution. If we know the median  $\mu$ , the only unknown in the 75th percentile is  $\sigma$ , so two probability statements about the 75th percentile can be translated into two probability statements about  $\sigma$ . In fact, our expert was far more comfortable thinking about  $T$  than about  $\log(T)$ , but fortunately, the 75th percentile of  $T$  is just  $e^{\mu+z_{0.75}\sigma} = Me^{z_{0.75}\sigma} = 1.963M$ .

Finally, to learn about  $\sigma_\delta$ , we elicit information about the difference between the best Ache hunter and an average Ache hunter. The  $\delta_i$ s are individual hunter effects. Rewriting equation (1) gives

$$T_{ij} = e^\mu e^{\delta_i} e^{\varepsilon_{ij}},$$

so  $e^{\delta_i}$  is a multiplicative effect for the  $i$ th hunter. Note that for an average hunter with  $\delta_i = 0$ , the multiplier is 1. A typical hunter effect is  $e^\delta$ .

We interpret the best hunter to be the one at the 90th percentile of all hunter effects  $e^\delta$ . This was our interpretation but was reconfirmed as reasonable by our expert. From (2), the 90th percentile of  $\delta$  is  $z_{0.90}\sigma_\delta = 1.282\sigma_\delta$ , so the 90th percentile of hunter effects is  $e^{z_{0.90}\sigma_\delta} \equiv Q$ . We elicited information about  $Q$ , by asking our expert to make probability statements about how much better the best hunter would be than an average hunter.

Dr. McMillan provided a best guess of 90 minutes for the median hunting time and was 95% sure that this median was under  $2\frac{1}{2}$  hours (150 minutes). With the median search time  $M = e^\mu$ , these statements specify a normal prior for  $\mu$  by solving the simultaneous equations

$$0.5 = \Pr(e^\mu \leq 90) = \Pr[\mu \leq \log(90)]$$

and

$$0.95 = \Pr(e^\mu \leq 150) = \Pr[\mu \leq \log(150)],$$

yielding  $\mu \sim N(4.5, 0.1)$ .

Our prior for  $\sigma$  is found by eliciting information on the random third quartile (75th percentile) of average Ache hunter search times *given a specified median search time*,  $Me^{\sigma z_{0.75}}$  with  $z_{0.75} = 0.6745$ . Given that  $M = 90$ , the expert provided a best guess of 180 minutes for the third quartile and was 95% sure that the third quartile is below 300 minutes. We translate these statements into

$$0.5 = \Pr(Me^{\sigma z_{0.75}} \leq 180 | M = 90) = \Pr(\sigma \leq 1.0276)$$

and

$$0.95 = \Pr(Me^{\sigma z_{0.75}} \leq 300 | M = 90) = \Pr(\sigma \leq 1.785).$$

The gamma prior  $\sigma \sim \text{Gamma}(7.602, 7.076)$  satisfies these conditions.

A randomly selected Ache male has a random *acceleration factor*  $e^\delta$  relative to Ache hunters of average skill (who have  $\delta = 0$ ). The expert believes that the best Ache hunters are about 1.5 times faster at finding burrows than average Ache hunters and at the most twice as fast. We assume that the best Ache hunters are those at  $Q$ , the upper 10% of hunting ability defined by the relationship  $\Pr(e^\delta \leq Q | \sigma_\delta) = 0.9$ . Then  $\Pr[\delta \leq \log(Q) | \sigma_\delta] = 0.9$ ,  $\log(Q) = z_{0.9}\sigma_\delta = 1.282\sigma_\delta$ , and  $Q = e^{z_{0.9}\sigma_\delta} = e^{1.282\sigma_\delta}$ . Approximating the expert's opinion, we have

$$\Pr(Q \leq 1.5) = 0.5 \quad \text{and} \quad \Pr(Q \leq 2) = 0.95,$$

so

$$P\left(\sigma_\delta \leq \frac{\log(1.5)}{1.282}\right) = 0.5 \quad \text{and} \quad P\left(\sigma_\delta \leq \frac{\log(2)}{1.282}\right) = 0.95.$$

We fit a gamma prior to  $\sigma_\delta$  by solving the above equations to obtain  $\sigma_\delta \sim \text{Gamma}(8.1, 24.5)$ .

---

## Chapter 2

---

# Fundamental Ideas I

---

There is a vast literature on Bayesian statistics. Four foundational works are de Finetti (1974 and 1975), Jeffreys (1961), and Savage (1954). Good elementary introductions to the subject are Lindley (1971) and Berry (1996). Early efforts to make Bayesian methods accessible for data analysis were made by Raiffa and Schlaifer (1961), Zellner (1971), and Box and Tiao (1973). The important topic of Bayesian prediction was presented in Aitchison and Dunsmore (1975) and Geisser (1993). Bayesian decision theory and more theoretical aspects of Bayesian inference were presented in DeGroot (1970) and Berger (1993). Modern Bayesian data analysis methods based on Markov chain Monte Carlo methods are presented in Gelman et al. (1995), Carlin and Louis (2008), Congdon (2001), and Marin and Robert (2007). Recent theoretical treatments are found in Robert (2007) and Bernardo and Smith (2000). The treatment presented here owes debts to many of these works.

Although Bayesian methodology allows every data analyst their own prior distribution, we believe that it remains consistent with the practice of science. For large amounts of data, scientists with different prior beliefs should ultimately agree after (separately) combining the data with their prior information. At least, this should happen for anyone with a “reasonable” prior. On the other hand, insufficient data can result in (continued) discrepancies of opinion about relevant scientific questions. In the real world, that is how science works. More philosophically, Bayesian statistics appears to be the only logically consistent method of making statistical inferences, although not the only useful one. Our presentation incorporates some non-Bayesian ideas, especially as they relate to model checking.

As seen in Chapter 1, we cover a wide array of examples. It was not always possible for us to use examples for which we had access to expert opinion. In some cases, we have specified our own personal priors. Other times we have used some version of a reference prior. In problems involving many parameters, it is rarely possible to elicit good prior information about all of them. We discuss partial prior information as a possibility in such cases. Our favorite examples are the ones where we got prior information by laboriously extracting it from experts.

Throughout we try to give applications precedence over mathematics. It is our intention to focus on statistical ideas, models, and interpretations while keeping the mathematical level as low as possible. Nonetheless, understanding the concepts of Bayesian analysis requires some understanding of multivariable calculus, especially multiple integration. Typical problems involve multiple observations and often involve multiple parameters. Bayesian statistics presumes the existence of a joint probability distribution over all these quantities and we must be prepared to find integrals computed over any or all of these quantities. Rather than using integral calculus, computer generated numerical approximations are typically used, even when the calculus is tractable.

Appendices A and B contain background material. In dealing with regression problems, we assume that the reader has some familiarity with basic matrix operations such as addition, multiplication, and the concepts of a transpose and an inverse matrix. We do not perform many sophisticated matrix manipulations, but for those unfamiliar with the subject a brief introduction is given in Appendix A. To perform Bayesian analysis, *previous exposure to probability theory is necessary*, including the idea of joint distributions. Probability concepts are briefly reviewed in Appendix B.

Computations are largely performed in WinBUGS but some are performed in the computing language R. Both are free on the Internet. Chapter 3 introduces WinBUGS and R. Appendix C examines some salient aspects of R.

For most purposes, we will not need to distinguish between probability density functions and probability mass functions, so we follow the “deplorable” convention of calling both density functions, cf. Cox (2006, Chapter 1). If necessary we distinguish between discrete and continuous cases by specifying discrete densities or continuous densities.

Throughout the book, sections with a higher level of mathematical sophistication are starred.

## 2.1 Simple Probability Computations

For two events  $A$  and  $B$  the conditional probability of  $A$  given  $B$  is defined as

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}.$$

$A \cap B$  denotes the intersection of  $A$  and  $B$ , i.e., the event that both  $A$  and  $B$  occur. Let  $A^c$  denote the complement of  $A$ , that is, all the outcomes that are not part of  $A$ .

Bayes’ Theorem allows us to compute  $\Pr(A|B)$  from  $\Pr(B|A)$ ,  $\Pr(B|A^c)$ , and  $\Pr(A)$  via

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B|A)\Pr(A) + \Pr(B|A^c)\Pr(A^c)}.$$

The theorem is the direct result of two facts: first, in the numerator, from the definition of conditional probability  $\Pr(B|A)\Pr(A) = \Pr(A \cap B)$  and, second, using the Law of Total Probability in the denominator, we have

$$\begin{aligned} \Pr(B) &= \Pr([A \cap B] \text{ or } [A^c \cap B]) = \Pr([A \cap B]) + \Pr([A^c \cap B]) \\ &= \Pr(B|A)\Pr(A) + \Pr(B|A^c)\Pr(A^c). \end{aligned}$$

**EXAMPLE 2.1.1. Drug Testing.** Let  $D$  indicate a drug user and  $C$  indicate someone who is clean of drugs. Let  $+$  indicate that someone tests positive for a drug, and  $-$  indicates testing negative. The overall *prevalence* of drug use in the population is  $\Pr(D) = 0.01$ , so in this population drug use is relatively rare. The *sensitivity* of the drug test is, say,  $\Pr(+|D) = 0.98$ . This is how good the test is at identifying people who use drugs. The *specificity* of the drug test is  $\Pr(-|C) = 0.95$ , which is how good the test is at correctly identifying nonusers. We want to find the probability that someone uses drugs given that they tested positive. From Bayes’ Theorem, the probability is

$$\begin{aligned} \Pr(D|+) &= \frac{\Pr(+|D)\Pr(D)}{\Pr(+|D)\Pr(D) + \Pr(+|C)\Pr(C)} \\ &= \frac{0.98 \times 0.01}{[0.98 \times 0.01] + [(1 - 0.95) \times (1 - 0.01)]} \\ &= \frac{0.98 \times 0.01}{[0.98 \times 0.01] + [0.05 \times 0.99]} \\ &= \frac{0.0098}{0.0098 + 0.0495} \\ &\doteq 0.165. \end{aligned}$$

Even after testing positive for drug use using a very good test, there is an 83% chance,  $1 - 0.165$ , that the person does not use drugs. The overwhelming probability is that a positive test outcome is false. This result is driven by the very low initial probability of drug use, so even after incorporating the positive test, the probability remains low. After testing positive, the probability of drug use is more

than 16 times greater than before the test, but it is still relatively low. This illustrates the difficulty in screening large numbers of people, most of whom are unlikely to be users. Most positive tests will be false positives. On the other hand, if the person was tested because their behavior indicated a 50/50 chance of drug use, the calculation becomes

$$\Pr(D|+) = \frac{0.98 \times 0.5}{[0.98 \times 0.5] + [0.05 \times 0.5]} = \frac{0.98}{0.98 + 0.05} \doteq 0.95$$

and the person is very likely to be a user.

Later, in Chapter 14 on binary diagnostic tests, we will revisit this problem when we do not actually know the prevalence, sensitivity, and specificity but rather have expert information about their values and data on the number of people who test positive.

**EXERCISE 2.1.** An enzyme-linked immunosorbent assay (ELISA) test is performed to determine if the human immunodeficiency virus (HIV) is present in the blood of individuals. The ELISA test is not perfect. Suppose that the ELISA test correctly indicates HIV 99% of the time, and that the proportion of the time that it correctly indicates no HIV is 99.5%. Suppose that the prevalence among blood donors is known to be 1/10,000. What proportion of blood that is donated will test positive using the ELISA test? Also, what proportion of the blood that tests negative on the ELISA test is actually infected with HIV? Finally, what is the probability that a positive ELISA outcome is truly positive, that is, what proportion of individuals with positive outcomes are actually infected with HIV?

**EXAMPLE 2.1.2. *Prosecutor's Fallacy.*** Suppose that the defendant in a criminal trial is either  $G$ , guilty, or  $I$ , innocent. In addition, suppose blood  $B$  has been found at the scene of the crime that is consistent with the defendant's. The prosecutor's fallacy is that if only 1% of the population has this blood type, then the probability that this defendant is innocent is 1%. In mathematical notation, *the fallacy* is that if  $\Pr(B|I) = 0.01$ , then  $\Pr(I|B) = 0.01$ . Surprisingly, this fallacious conclusion is not always as ridiculous as you might think, but it is inconsistent with the presumption of innocence.

We compute  $\Pr(G|B) = 1 - \Pr(I|B)$ , which the fallacy says should be 0.99. The key fact to recognize is that if the defendant is guilty, the blood type will certainly match, i.e.,  $\Pr(B|G) = 1$ . Applying Bayes' Theorem and the fact that  $\Pr(I) = 1 - \Pr(G)$ ,

$$\begin{aligned}\Pr(G|B) &= \frac{\Pr(B|G)\Pr(G)}{\Pr(B|G)\Pr(G) + \Pr(B|I)\Pr(I)} \\ &= \frac{1 \times \Pr(G)}{[1 \times \Pr(G)] + [0.01 \times \Pr(I)]} \\ &= \frac{\Pr(G)}{\Pr(G) + 0.01[1 - \Pr(G)]} \\ &= \frac{\Pr(G)}{0.99\Pr(G) + 0.01} \\ &= \frac{1}{0.99 + 0.01/\Pr(G)}.\end{aligned}$$

If  $\Pr(G) = 1/2$ , then  $\Pr(G|B) = 1/1.01 \doteq 0.99$ , so the “prosecutor's fallacy” is approximately the appropriate posterior probability *if the defendant has a 50% prior chance of being guilty*. Of course this is hardly consistent with the presumption that defendants are innocent, but may be more generous towards the defendant than what typically occurs in court. One gets the impression that juries tend to assume that defendants are guilty. Note also that  $\Pr(G) = 0.01$  implies  $\Pr(G|B) \doteq 0.5$ , so a presumption of innocence changes the problem dramatically. Of course, if one assumes the person is guilty, data will not change that presumption,  $\Pr(G) = 1$  implies  $\Pr(G|B) = 1$ .

The prosecutor's fallacy generally works as illustrated for small values of  $\Pr(B|I)$ . For example, if  $\Pr(B|I) = 0.1$  and  $\Pr(G) = 0.5$ , then  $\Pr(G|B) = 0.91 \doteq (1 - 0.1)$ . On the other hand, if  $\Pr(B|I) = 0.5$  and  $\Pr(G) = 0.5$ , then  $\Pr(G|B) = 2/3$ , which is not well approximated by  $0.5 = (1 - 0.5)$ . See [buchanan.blogs.nytimes.com/2007/05/16/the-prosecutors-fallacy/](http://buchanan.blogs.nytimes.com/2007/05/16/the-prosecutors-fallacy/) for a real life application of the prosecutor's fallacy.

**EXAMPLE 2.1.3. *Defendant's Fallacy.*** We use much the same notation as in the previous example. The defendant's fallacy is that if there are 500,000 people in a town, and 1% have the same blood type as the defendant, then there are  $(500,000)(0.01) = 5,000$  people in town who have the blood type, so the probability of the defendant being guilty should be  $\Pr(G|B) = 1/5,000$ . In probabilistic notation, *the defendant's fallacy* is  $\Pr(G|B) = \Pr(G)/\Pr(B|I)$ , where  $\Pr(G) = 1/500,000$  and  $\Pr(B|I) = 0.01$ . Again, this can be a surprisingly reasonable conclusion. As before,  $\Pr(B|G) = 1$ , but now we are making a presumption of innocence in the form of  $\Pr(G) = 1/500,000$ . Bayes' Theorem gives

$$\begin{aligned}\Pr(G|B) &= \frac{\Pr(B|G)\Pr(G)}{\Pr(B|G)\Pr(G) + \Pr(B|I)\Pr(I)} \\ &= \frac{1 \left( \frac{1}{500,000} \right)}{1 \left( \frac{1}{500,000} \right) + 0.01 \left( \frac{499,999}{500,000} \right)} \\ &\doteq \frac{1}{5,000}.\end{aligned}$$

While the prosecutor's fallacy breaks down when  $\Pr(B|I)$  gets large, the defendant's fallacy breaks down when  $\Pr(B|I)$  gets small. For example, if  $\Pr(B|I) = 1/500,000$ , the defendant's fallacy suggests that  $\Pr(G|B)$  should be 1, whereas the posterior probability is only about 0.5.

Our next example involves more than two possible outcomes. In such cases, Bayes' Theorem becomes a little more complicated. Suppose  $E_1, \dots, E_k$  are distinct events that include all possible outcomes, that is, they are “mutually exclusive and exhaustive.” Let  $A$  be some other event. First, we can write the probability of  $A$  as

$$\Pr(A) = \Pr(A|E_1)\Pr(E_1) + \dots + \Pr(A|E_k)\Pr(E_k).$$

This is a version of the “Law of Total Probability.” We can now extend Bayes' Theorem to compute the probability of  $E_i$  given that  $A$  has occurred:

$$\Pr(E_i|A) = \frac{\Pr(A|E_i)\Pr(E_i)}{\Pr(A)} = \frac{\Pr(A|E_i)\Pr(E_i)}{\Pr(A|E_1)\Pr(E_1) + \dots + \Pr(A|E_k)\Pr(E_k)}.$$

**EXAMPLE 2.1.4. *The Monte Hall Problem – Or, Dealing While on the Make.*** On the antediluvian television show *Let's Make a Deal*, hosted by Monte Hall, the grand prize was awarded in the following manner. The prize was placed behind one of three doors. The contestant selected a door. Monte then showed the contestant what was behind one of the other two doors but it was never the grand prize. Finally, the contestant was allowed either to keep their initial choice or switch to the remaining unopened door. Some people's intuition is that there is a 50/50 chance that the prize is behind either of the two remaining unopened doors, so it would not matter if you switch. In fact, the probability is 2/3 that the prize is behind the other door that Monte did not open. One intuitive way to arrive at this conclusion argues that you already know the prize is not behind one of the two doors you did not select and the fact that Monte showed you it was not behind one of them gives you no additional information. However, by switching from your initial choice, essentially you are

being allowed to get both of the other two doors, and thus have a  $2/3$ s chance of getting the prize. This argument is rather inexact, so we now give a careful argument using Bayes' Theorem.

There are three variables involved. Let  $P$  denote the door that contains the prize, let  $C$  denote the door that you initially chose, and let  $S$  denote the door that Monte shows you. We assume that the prize is randomly placed behind a door so  $\Pr(P = p) = 1/3$  for  $p = 1, 2, 3$ . Also, the prize is placed prior to your choice of door, so it is independent of  $C$  and

$$1/3 = \Pr(P = p) = \Pr(P = p|C = c) \quad p = 1, 2, 3, c = 1, 2, 3.$$

What Monte shows you depends on where the prize is and what door you have chosen, so the door he shows you is selected according to a conditional probability  $\Pr(S = s|P = p, C = c)$ . For the sake of simplicity, we will assume that you initially chose door number 1, so throughout we have  $C = 1$ . All of our computations depend on  $C = 1$ , which will be implicit in what follows, so we write

$$\Pr(P = p) \equiv \Pr(P = p|C = 1),$$

etc., and for additional economy we write

$$f(s|p) \equiv \Pr(S = s|P = p, C = 1).$$

According to the rules of the game, Monte never shows you the prize. If the prize is behind door number 1 (the one we initially chose), Monte randomly picks either door 2 or door 3 and shows it to us. If the prize is behind door number 2, Monte shows us door 3. If the prize is behind door number 3, Monte shows us door 2. We summarize as follows:

$s$	1	2	3
$f(s 1)$	0	0.5	0.5
$f(s 2)$	0	0	1
$f(s 3)$	0	1	0.

Note that the column of 0s for  $s = 1$  is a result of our initial choice  $C = 1$ . It simply reflects that Monte never shows us what is behind our door. Now suppose that Monte shows us door number 2. We want to know the probabilities that the prize is behind door 1,  $\Pr(P = 1|S = 2)$ , and that the prize is behind door 3,  $\Pr(P = 3|S = 2)$ . We already know that  $\Pr(P = 2|S = 2) = 0$ . Using Bayes' Theorem

$$\begin{aligned} & \Pr(P = 1|S = 2) \\ &= \frac{f(2|1)\Pr(P = 1)}{[f(2|1)\Pr(P = 1)] + [f(2|2)\Pr(P = 2)] + [f(2|3)\Pr(P = 3)]} \\ &= \frac{0.5(1/3)}{[0.5(1/3)] + [0(1/3)] + [1(1/3)]} \\ &= 1/3 \end{aligned}$$

whereas

$$\begin{aligned} & \Pr(P = 3|S = 2) \\ &= \frac{f(2|3)\Pr(P = 3)}{[f(2|1)\Pr(P = 1)] + [f(2|2)\Pr(P = 2)] + [f(2|3)\Pr(P = 3)]} \\ &= \frac{1(1/3)}{[0.5(1/3)] + [0(1/3)] + [1(1/3)]} \\ &= 2/3, \end{aligned}$$

so as advertised, the probability of getting the prize is  $2/3$  if we switch doors. Similarly, if Monte shows us door number 3, we have  $\Pr(P = 1|S = 3) = 1/3$ ,  $\Pr(P = 2|S = 3) = 2/3$ , and  $\Pr(P = 3|S = 3) = 0$ . Similar results hold regardless of the initial choice  $C$ .

**EXERCISE 2.2.** The computation depends on Monte randomly picking between doors 2 and 3 when  $P = 1$  and  $C = 1$ . Suppose that whenever he was allowed to, Monte always showed door number 2, that is,  $f(2|1) = 1$ . Show that  $\Pr(P = 1|S = 2) = \Pr(P = 3|S = 2) = 0.5$  but also that  $\Pr(P = 1|S = 3) = 0$  and  $\Pr(P = 2|S = 3) = 1$ . Find the relevant probabilities when  $f(2|1) = 0.75$ .

**EXERCISE 2.3.** Automobiles are manufactured 5 days a week. Suppose one eighth of all cars are made on Mondays and Fridays, respectively, and that one fourth are made on Tuesdays, Wednesdays, and Thursdays. Also assume that on Mondays and Fridays, defective cars are produced at a rate of one out of every 20, while during the rest of the week they are produced at a rate of one out of every 100. What proportion of cars are defective during a typical week? If you just bought a defective car, what are the chances that it was manufactured on a Friday? A Tuesday?

**EXERCISE 2.4.** An article in the July 22, 2009 edition of the International Herald Tribune indicated that 7% of British children attend special schools that cater to privileged parents and that 75% of all judges are known to have attended such schools. The implication was that underprivileged, but presumably intelligent and hardworking children were being excluded from high ranking professions like judgeships. We want to look at the relative probabilities of being a judge given that one did or did not attend an elite school.

Specifically, let  $E$  denote attending an elite school with  $E^c$  the complement. Let  $J$  denote becoming a judge with  $J^c$  the complement. Let  $p = \Pr(J)$  be the (unknown to us) proportion of judges among the populace in Great Britain. We are given that  $\Pr(E|J) = 0.75$  and  $\Pr(E) = 0.07$ . (a) Use the definition of conditional probability to find  $\Pr(J|E)$  and  $\Pr(J|E^c)$  as functions of  $p$ . Find an actual number for the ratio  $\Pr(J|E)/\Pr(J|E^c)$ . What can you say about the effect of availability of elite schooling on the prospects of becoming a judge in Great Britain? (b) Let  $q = \Pr(E)$ . What value of  $q$  would correspond to no effect of  $E$  on the chances of becoming a judge later in life?

## 2.2 Science, Priors, and Prediction

Fundamentally, the field of statistics is about using probability models to analyze data. There are two major philosophical positions about the use of probability models. One is that probabilities are determined by the outside world. The other is that probabilities exist in people's heads. Historically, probability theory was developed to explain games of chance. For example, the physical structures involved in rolling dice, spinning roulette wheels, and dealing well-shuffled decks of cards suggest obvious probabilities for various outcomes. The notion of probability as a belief is more subtle. For example, suppose that you are in my presence when I flip a coin. Prior to flipping the coin, the physical mechanism involved suggests probabilities of 0.5 for each of the outcomes heads and tails. But now I have flipped the coin, looked at the result, but not told you the outcome. As long as you believe I am not cheating you, you would naturally continue to describe the probabilities for heads and tails as 0.5. But this probability is no longer the probability associated with the physical mechanism involved, because you and I have different probabilities. I know whether the coin is heads or tails, and your probability is simply describing your personal state of knowledge.

Bayesian statistics starts by using (*prior*) probabilities to describe your current state of knowledge. It then incorporates information through the collection of data. This results in new (*posterior*) probabilities to describe your state of knowledge after combining the prior probabilities with the data. *In Bayesian statistics, all uncertainty and all information are incorporated through the use of probability distributions, and all conclusions obey the laws of probability theory.*

Some form of prior information is always available. If we were trying to estimate mean height of American men, we all know that the mean does not exceed seven feet. The increased crop yield per acre treated with a new fertilizer is quite unlikely to exceed the weight of Greenland. At a minimum, a prior distribution for a parameter  $\theta$  can easily exclude values that are simply unrealistic and it

seems silly to ignore such information. Our goal is not to find the “perfect” prior distribution but rather to incorporate salient features of available scientific knowledge into the data analysis. As such, we need to examine whether other reasonable prior distributions lead to qualitatively similar posterior conclusions.

An agronomist plans to collect data to estimate the effect  $\theta$  of using a “new improved” fertilizer relative to a standard brand in growing corn. She expects results similar to experiments done by others but with variations due to differences in soil and climate. If an estimated effect of  $\hat{\theta} = 5$  bushels per acre had been presented in the literature with a 95% interval of two to eight bushels per acre, that information can be used as the basis for a prior distribution. Alternatively, as a tool in specifying a prior, the agronomist might be asked to provide a best guess for  $\theta$  and an interval in which she is, say, 95% certain that  $\theta$  would lie.

Unfortunately, there are many distributions with the same 95% interval. We could ask the agronomist to specify a number of percentiles of the distribution including the median, but still there would be multiple distributions with the same percentiles. Moreover, this process of eliciting information from an expert is surprisingly laborious for both the expert and the statistician.

Typically, from an expert we obtain some number of characteristics about the population under study, obtaining as much information as is reasonable and convenient to elicit. We then identify a prior from some suitable class of distributions that agrees with the characteristics. Ideally the characteristics are based on other data that were collected independently of the current data, but expert knowledge comes in many forms. Our goal is to incorporate such knowledge into a suitable probability distribution that reflects uncertainty about model parameters. *After statisticians develop such a prior distribution, they should always return to the expert to validate that the prior is a reasonable approximation to the expert’s actual information.*

For example, suppose our agronomist essentially gets direct observations on the differences in corn yields for the two fertilizers, as from a paired comparison design. The data are modeled as  $N(\theta, \sigma^2)$ , where  $\sigma$  is the standard deviation of a random difference in yields. Our agronomist thinks that 5 is a reasonable guess for the mean yield difference  $\theta$ , and has provided a 95% interval of (2, 8). If we use a normal distribution to model prior information about  $\theta$ , say,  $\theta \sim N(\theta_0, b^2)$ , then the 95% interval must have endpoints of about  $\theta_0 \pm 2b$ . Setting  $\theta_0 = 5$  and  $\theta_0 + 2b = 8$ , we find  $b = 1.5$ . The size of  $b$  is directly related to how narrowly the expert specifies the 95% interval. Note that this technique used the fact that (2, 8) is centered around  $\theta_0 = 5$ . To get a unique prior, the number of parameters in the class of priors must be the same as the number of distinct characteristics specified.

Often it is difficult for experts to provide direct information on the parameters of statistical models because they are not comfortable with the parameterization. In such cases, we elicit information about quantities that are more familiar to the expert and use that information to induce a prior on the parameters of the statistical model. Typically, agronomic experts would have good ideas about mean differences  $\theta$  due to a fertilizer and could give a best guess and a probability interval for it. It is more difficult for experts to think directly about likely values for the standard deviation  $\sigma$ . Instead, we can ask them to think about, say, the upper quartile of the data: the number that has 75% of the differences below it and 25% above. For our normal distribution with mean  $\theta$ , the upper quartile is  $\theta + 0.675\sigma$ , so if our expert’s experience suggested that the upper quartile was about 10, then we can use their best guess for  $\theta$  of  $\theta_0 = 5$  and obtain a best guess for  $\sigma$  of  $\sigma_0 = (10 - 5)/0.675 = 7.41$ . For  $\sigma^2$ , we would pick a distribution that had a median or mode equal to the best guess,  $(7.41)^2$ . We defer detailed discussion of priors for  $\sigma^2$  until Chapter 5.

Although parameters are often mere conveniences, frequently the parameter  $\theta$  has some basis in physical reality. Rather than describing where  $\theta$  really is, the prior describes beliefs about where  $\theta$  is. The probabilities specified by the agronomist are her beliefs about  $\theta$ . Different agronomists, even those working on the same crops in the same region, will have different knowledge bases and therefore different probabilities for  $\theta$ . If they analyze the same data, they will continue to have different opinions about  $\theta$  until sufficient data are collected so that their beliefs converge and a

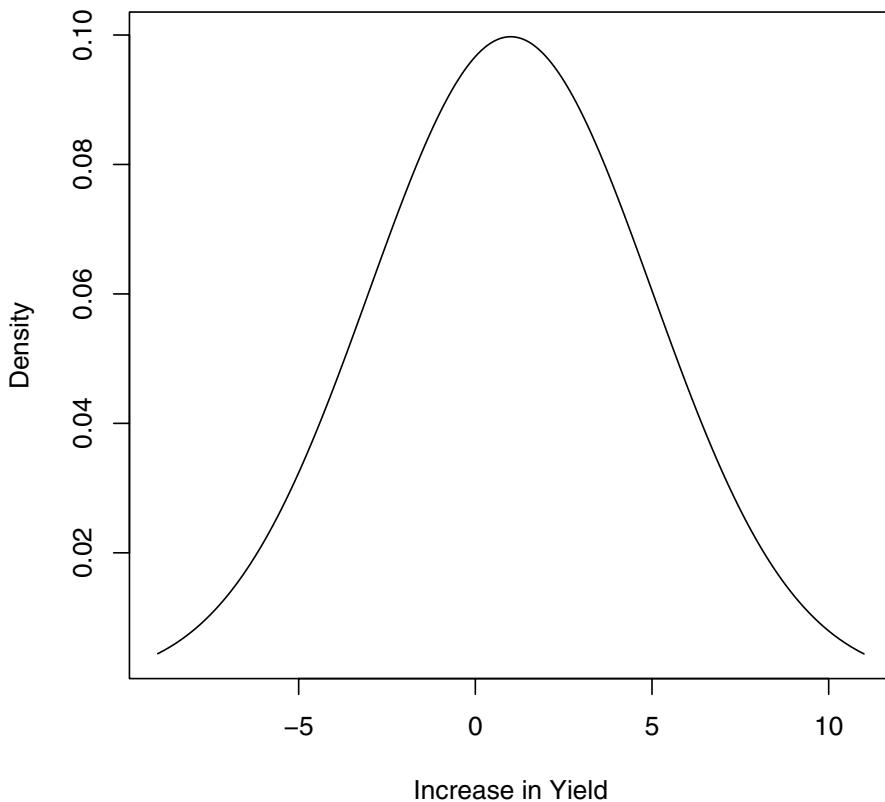


Figure 2.1: *Predictive density for future difference in yields with  $\theta = 1$ .*

consensus is reached. This should occur unless one or more of them is unrealistically dogmatic. For example, when considering ESP, some people refuse to place any positive probability on the phenomenon's existence, so no amount of data will change their beliefs, cf. Utts (1991). Of course, Bayesian statistics is primarily a tool for evaluating relative evidence. It is not well suited for falsifying an individual hypothesis, cf. Section 4.1. For an excellent reference on the historical role of subjectivity in science, see Press and Tanur (2001).

In the corn yield example, individual farmers will have less interest in  $\theta$  than in how the new fertilizer will change *their* yields. Suppose the (predictive) distribution for changes in yields is normal with mean  $\theta = 1$  bushel per acre and a standard deviation of 4 as illustrated in Figure 2.1. The probability of an improvement of at least one bushel per acre is 0.5, the area under this curve to the right of 1. We similarly note that the area to the left of 0 is about 0.4. So there is a 40% chance of having a worse yield under the new fertilizer. While the “typical” farmer expects a 1 bushel per acre increase, half the farmers will do worse than this and 40% will actually see yields decline. Taking into account the risks and benefits of changing yields and the costs of switching fertilizers, the change may or may not be worth the risk.

Consider Figure 2.2, which gives another predictive density, but now with a mean of  $\theta = 3$  bushels per acre and a smaller standard deviation 2. There is now a 93% chance that the difference in yields will be positive, so the farmer might be much more inclined to switch fertilizers. However,

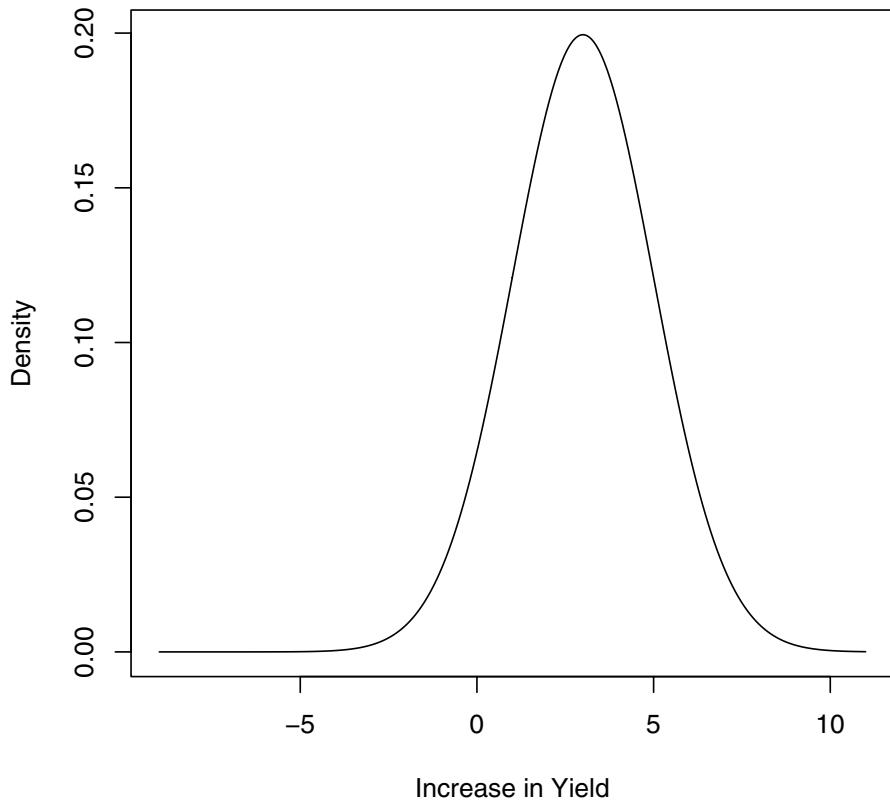


Figure 2.2: Predictive density for future difference in yields with  $\theta = 3$ .

in Iowa 200 bushels per acre is common, so improvements of 1 or 3 bushels per acre might be unimportant to Iowa farmers. On the other hand, if you have 1,000 acres in corn, an additional 1,000 to 3,000 bushels might be worth some trouble to obtain. It's all relative. The farmer needs good information and a relevant predictive distribution is far more useful than any point or interval estimate for  $\theta$ .

Rather than getting a predictive distribution for the change in yields, the farmer may want individual predictive distributions for yields based on the standard treatment and on the new treatment. Figure 2.3 gives predictive densities for individual yields under fertilizers 3 and 4 respectively. Fertilizer 3 has a mean yield of 200 bushels per acre and fertilizer 4, with the shaded curve, has a mean yield of 220 bushels per acre. If buying that new combine cannot happen unless our farmer gets at least 250 bushels per acre, the probability under fertilizer 3 is only 0.006 while it is 0.27 under fertilizer 4. On the other hand, if a yield below 150 bushels per acre means bankruptcy, there is only probability 0.006 under fertilizer 3, while there is probability 0.08, nearly 10%, of seeing a yield that low under fertilizer 4. The "risk taker" farmer might switch to fertilizer 4, betting that she will be one of the lucky ones with a yield above 250 but understanding that she could come in below 150. A risk averse farmer might stay with fertilizer 3 and be virtually guaranteed of both staying in business and having to delay purchase of a combine.

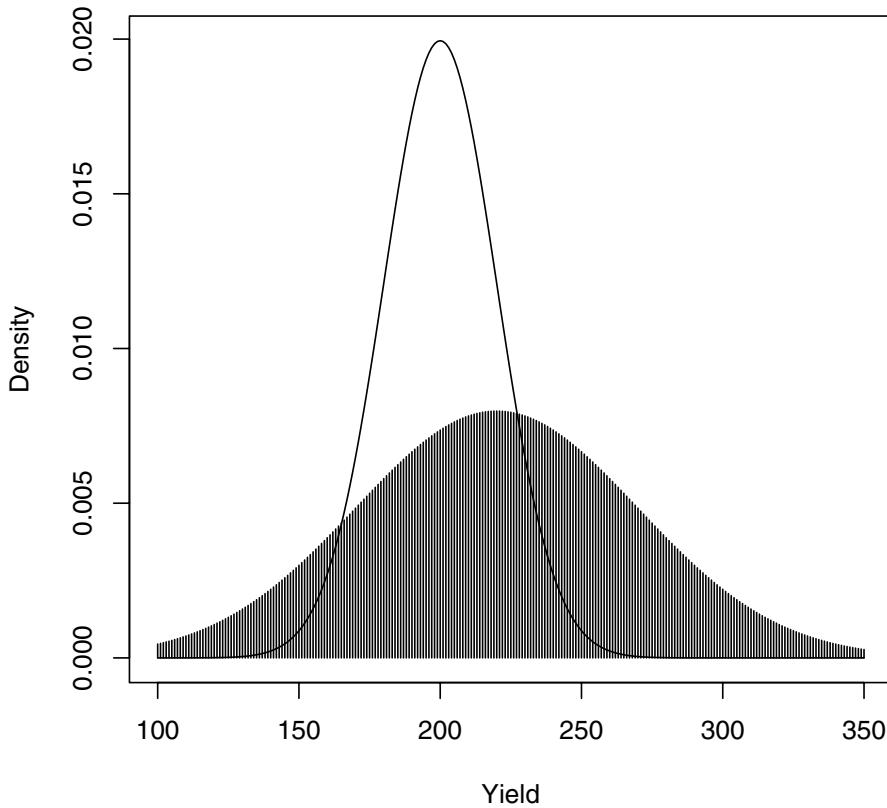


Figure 2.3: *Predictive densities for future yields under two fertilizers. Shaded curve (fertilizer 4) has mean yield of 220 while unshaded curve (fertilizer 3) has mean yield of 200.*

### 2.3 Statistical Models

Science is fundamentally about taking observations and predicting new observations. Statistical models are useful tools for scientific prediction. The parameters  $\theta$  are often selected for convenience in building models that will predict well. The use of parameters is not a fundamental aspect of Bayesian analysis. A wise man once wrote that parameters are “too often created by and for statisticians so that they have something to draw inferences about,” Christensen (2002, p. 131). (Ok, maybe he was a wiseguy and not a wise man.) Nonetheless, there are situations where parameters are so closely related to the behavior of observable quantities that the parameters become of interest in their own right. For example, the parameter  $\theta$  might be the population mean blood pressure of people taking a particular drug in a randomized trial. It is not a great leap from seeing that  $\theta$  is relatively small to concluding that the drug may help reduce the blood pressure observations of people who take it. However, in many other examples, the relationship between parameters and observations is not nearly so obvious. It is our goal to focus on observables (prediction) and parameters that are closely related to observables. (But goals are often imperfectly achieved.) Before discussing prediction, we discuss posterior distributions for the parameters of a statistical model.

Statistical models typically involve multiple observations (random variables), say,  $y_1, \dots, y_n$ . Dealing with these is facilitated by writing them collectively as a vector of observations, say

$y = (y_1, \dots, y_n)'$  where the  $'$  indicates transposing of the row vector so that  $y$  is an  $n \times 1$  matrix, see Appendix A. Typically, the observations are collected independently given the parameters of the model. Denote the parameters  $\theta = (\theta_1, \dots, \theta_r)'$ . In many simple problems  $r = 1$ .

Bayesian statistics typically begins with prior information about the *state of nature*  $\theta$  that is embodied in the prior density  $p(\theta)$ . It then uses Bayes' Theorem and the random data  $y$ , with sampling density  $f(y|\theta)$ , to update this information into a posterior density  $p(\theta|y)$  that incorporates both the prior information and the data. Specifically, *Bayes' Theorem* tells us that

$$p(\theta|y) = \frac{f(y|\theta)p(\theta)}{\int f(y|\theta)p(\theta)d\theta}.$$

We illustrate these ideas with Binomial, Bernoulli, and normal data. These are examples in which the mathematics is relatively simple but by no means trivial. We include the mathematics here to give a better understanding of what we try to avoid in the remainder of the book. Throughout we use the *indicator function* of a set  $A$  defined as

$$I_A(\theta) = \begin{cases} 1 & \text{if } \theta \in A \\ 0 & \text{if } \theta \notin A \end{cases}.$$

Occasionally, this gets written  $I(\theta \in A)$  where the function is 1 if the logical expression  $\theta \in A$  is true and the function is 0 if the logical expression is not true. Similar indicators are defined when  $\theta \in A$  is replaced by other logical expressions.

**EXAMPLE 2.3.1.** *Binomial Data.* Suppose we are interested in assessing the proportion of U.S. transportation industry workers who use drugs on the job. Let  $\theta$  denote this proportion and assume that a random sample of  $n$  workers is to be taken while they are actually on the job. Each individual will be strapped down and forced to deliver a blood sample which will be analyzed for a variety of legal and illegal drugs. (We certainly would not want our bus drivers on steroids, they are already cranky enough.) The number of positive tests is denoted  $y$ . In particular, we took  $n = 10$  samples and obtained  $y = 2$  positive test results.

Data obtained like this follow a Binomial distribution. Write

$$y|\theta \sim \text{Bin}(n, \theta),$$

with discrete density function for  $y = 0, 1, \dots, n$

$$f(y|\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}.$$

It will be convenient, but by no means necessary, to use a Beta distribution as a model for the prior. The Beta distribution is conjugate to the Binomial distribution in the sense that the densities have similar functional forms. Let

$$\theta \sim \text{Beta}(a, b)$$

with density

$$p(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1} I_{(0,1)}(\theta). \quad (1)$$

Here  $\Gamma(\cdot)$  is the well-known Gamma function and, as will be discussed soon, the *hyperparameters*  $a$  and  $b$  are selected to reflect the researcher's beliefs and uncertainty.

The posterior distribution turns out to be another Beta distribution, in particular the distribution of  $\theta$  given  $y$  is

$$\theta|y \sim \text{Beta}(y+a, n-y+b).$$

In our transportation example we observed  $n = 10$  and  $y = 2$ . As justified later, we use a prior with  $a = 3.44$  and  $b = 22.99$ , so the posterior is

$$\theta|y \sim \text{Beta}(2 + 3.44, 10 - 2 + 22.99) = \text{Beta}(5.44, 30.99).$$

To obtain the posterior, apply Bayes' Theorem to the densities

$$\begin{aligned} p(\theta|y) &= \frac{f(y|\theta)p(\theta)}{\int f(y|\theta)p(\theta)d\theta} \\ &= \frac{\binom{n}{y}\theta^y(1-\theta)^{n-y}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}I_{(0,1)}(\theta)}{\int_0^1 \binom{n}{y}\theta^y(1-\theta)^{n-y}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}d\theta} \\ &= \frac{\theta^y(1-\theta)^{n-y}\theta^{a-1}(1-\theta)^{b-1}I_{(0,1)}(\theta)}{\int_0^1 \theta^y(1-\theta)^{n-y}\theta^{a-1}(1-\theta)^{b-1}d\theta} \\ &= \frac{\theta^{y+a-1}(1-\theta)^{n-y+b-1}I_{(0,1)}(\theta)}{\int_0^1 \theta^{y+a-1}(1-\theta)^{n-y+b-1}d\theta} \\ &= \frac{\Gamma(n+a+b)}{\Gamma(y+a)\Gamma(n-y+b)}\theta^{y+a-1}(1-\theta)^{n-y+b-1}I_{(0,1)}(\theta). \end{aligned}$$

The last equality is the trickiest and we discuss it next but note that by comparison with (1) this is the density of a  $\text{Beta}(y+a, n-y+b)$  distribution.

To obtain the last equality of the posterior density derivation, recall that any density must integrate to 1, so for a  $\text{Beta}(a, b)$

$$1 = \int_0^1 \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}d\theta.$$

It follows immediately that

$$\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \int_0^1 \theta^{a-1}(1-\theta)^{b-1}d\theta,$$

and correspondingly

$$\frac{\Gamma(y+a)\Gamma(n-y+b)}{\Gamma(n+a+b)} = \int_0^1 \theta^{y+a-1}(1-\theta)^{n-y+b-1}d\theta,$$

from which the last equality of the posterior density derivation follows.

In the posterior  $\theta|y \sim \text{Beta}(y+a, n-y+b)$ , the number of “successes”  $y$  and the hyperparameter from the prior  $a$  play similar roles. Also, the number of “failures”  $n-y$  and  $b$  play similar roles. We can think of the prior as augmenting the data with  $a$  successes and  $b$  failures out of  $a+b$  trials. Priors that allow such an interpretation are called *data augmentation priors (DAPs)*. In our transportation example with  $n = 10$  we have used a rather potent prior with  $a+b = 3.44 + 22.99 = 26.43$  prior observations.

In DAPs, the prior density  $p(\theta)$  has the same functional form as the sampling density  $f(y|\theta)$  when viewed as a function of  $\theta$ . That is to say,  $f(y|\theta)p(\theta) \propto f(y_*|\theta)$  for some new vector  $y_*$  and every vector  $\theta$ . However,  $y_*$  may not be subject to the same restrictions as  $y$ , e.g., for binomial data  $y$  and  $n$  are integers but  $a$  and  $b$  need not be.

Once we have the density of the posterior distribution, there are many things we might choose to use as summaries of our information. Both the prior and the posterior are Beta distributions so to simplify notation, we list some of these summaries for the prior distribution. The prior mean is

$$E(\theta) = \frac{a}{a+b} \equiv \mu.$$

We can write the prior variance in terms of  $a$  and  $b$  or in terms of the prior mean  $\mu$  and the prior sample size  $\psi \equiv a+b$ :

$$\text{Var}(\theta) = \mu(1-\mu)/(\psi+1).$$

The prior mode is

$$\frac{a-1}{\psi-2},$$

provided  $a > 1, b > 1$ . Many other items of interest cannot be computed analytically, so in Chapter 3 we discuss computer simulations of them. These include the median and the probability that  $\theta$  falls into any particular set, e.g.,  $\Pr[\theta > 0.5]$ .

The posterior is a Beta(5.44, 30.99) distribution, so for our data

$$E(\theta|y) = \frac{y+a}{n+a+b} = \frac{5.44}{36.43} = 0.15.$$

Interestingly, with the DAP prior, the posterior mean can be written as a weighted average of the data proportion of successes and the prior proportion of successes with the weights being the relative sizes of the actual data set and the prior data set, that is,

$$E(\theta|y) = \left(\frac{n}{\psi+n}\right)\left(\frac{y}{n}\right) + \left(\frac{\psi}{\psi+n}\right)\mu.$$

Now consider prediction. Suppose we are to observe a future binomial value,  $\tilde{y}$ , assumed to be a  $\text{Bin}(m, \theta)$  random variable that is independent of  $y$  given  $\theta$ . The predictive density is

$$f(\tilde{y}|y) = \int_0^1 f(\tilde{y}|y, \theta)p(\theta|y)d\theta = \int_0^1 f(\tilde{y}|\theta)p(\theta|y)d\theta,$$

where the second equality is due to the conditional independence of  $\tilde{y}$  and  $y$  given  $\theta$ . With

$$f(\tilde{y}|\theta) = \binom{m}{\tilde{y}}\theta^{\tilde{y}}(1-\theta)^{m-\tilde{y}},$$

we obtain for  $\tilde{y} = 0, \dots, m$ ,

$$\begin{aligned} f(\tilde{y}|y) &= \binom{m}{\tilde{y}} \frac{\Gamma(a+b+n)}{\Gamma(a+y)\Gamma(b+n-y)} \int_0^1 \theta^{a+y+\tilde{y}-1} (1-\theta)^{b+n-y+m-\tilde{y}-1} d\theta \\ &= \binom{m}{\tilde{y}} \frac{\Gamma(a+b+n)}{\Gamma(a+y)\Gamma(b+n-y)} \frac{\Gamma(a+y+\tilde{y})\Gamma(b+n+m-y-\tilde{y})}{\Gamma(a+b+n+m)}. \end{aligned}$$

This is called a *Beta-Binomial distribution*. Our main point is that there is an analytic expression for the predictive density and there are some other analytic results available but they are limited both in number and utility. For example, it is not difficult to see using iterated expectations as discussed in Proposition B.1 that, as might be anticipated,

$$E(\tilde{y}|y) = E_{\theta|y}[E(\tilde{y}|y, \theta)] = E_{\theta|y}[E(\tilde{y}|\theta)] = E_{\theta|y}[m\theta] = mE(\theta|y).$$

The posterior mean of  $\theta$  was given earlier. Similarly, the predictive variance is obtained analytically using an iterated variance formula. But even to compute probabilities involves using the gamma function which must often be evaluated numerically. Simulating the predictive distribution will give us much more flexibility in providing useful statistical analysis.

**EXERCISE 2.5.** Let  $\mu = 15$  and  $y/n = 10$  in the Binomial example. Then define  $n/\psi = k$  to be the ratio of the data sample size to the “prior sample size.” Plot  $E(\theta|y)$  versus  $k$  for  $k \in \{10, 5, 1, 1/5, 1/10\}$ .

In the Binomial example, the form of the sampling density  $f(y|\theta)$  was immediate. More often, we have to work to obtain it. The simplest case is when the data  $y_1, \dots, y_n$  are *independent and identically distributed (iid)*. In this case we can easily build  $f(y|\theta)$  from the density of an individual observation, say,  $f_*(y_i|\theta)$ . The data are identically distributed, so  $f_*$  applies to all observations, and because they are independent, their joint density is the product of their individual densities, so

$$f(y|\theta) = \prod_{i=1}^n f_*(y_i|\theta).$$

**EXAMPLE 2.3.2. Bernoulli Data.** Again suppose we are interested in assessing the proportion of U.S. transportation industry workers who use drugs on the job and  $\theta$  denotes this proportion. We take independent observations on individual workers. Let  $y_i$  be a one if the  $i$ th individual tests positive for drugs and zero otherwise.

Technically,  $\theta$  is the probability that someone in the population would have tested positive for drugs if they had the bad luck to be in our sample. We presume that this is a reasonable definition for the proportion of workers who use drugs on the job. Our statistical model is that  $y_i = 1$  with probability  $\theta$  and  $y_i = 0$  with probability  $1 - \theta$ . By definition, each  $y_i$  is called a *Bernoulli random variable* with parameter  $\theta$ . Given  $\theta$ , we assume the  $y_i$ s are independent, so  $y_1, \dots, y_n$  are iid Bernoulli with parameter  $\theta$ . Another way to write this is

$$y_1, \dots, y_n | \theta \stackrel{iid}{\sim} \text{Bern}(\theta).$$

The density for an individual  $y_i$  is a  $\text{Bin}(1, \theta)$  density,

$$f_*(y_i|\theta) = \theta^{y_i} (1 - \theta)^{1-y_i}, \quad y_i = 0, 1,$$

which simplifies to the probability of success,  $\theta$ , when we have a success, i.e.,  $y_i = 1$ , and the probability of failure,  $(1 - \theta)$ , when we have a failure,  $y_i = 0$ . Because the  $y_i$ s are iid, the *sampling density* of  $y = (y_1, \dots, y_n)'$  is

$$f(y|\theta) = \prod_{i=1}^n f_*(y_i|\theta) = \prod_{i=1}^n \theta^{y_i} (1 - \theta)^{1-y_i} = \theta^{\sum_{i=1}^n y_i} (1 - \theta)^{n - \sum_{i=1}^n y_i}. \quad (2)$$

As with the Binomial data, it is convenient to choose

$$\theta \sim \text{Beta}(a, b)$$

with  $a = 3.44$  and  $b = 22.99$ . As shown later, the posterior distribution turns out to be

$$\theta|y \sim \text{Beta}\left(\sum y_i + a, n - \sum y_i + b\right).$$

In our transportation example we observed  $n = 10$  workers and the second and seventh had positive tests, i.e.,  $y = (0, 1, 0, 0, 0, 0, 1, 0, 0, 0)'$  so that  $\sum y_i = 2$  and  $n - \sum y_i = 8$ . The posterior is

$$\theta|y \sim \text{Beta}(2 + 3.44, 10 - 2 + 22.99) = \text{Beta}(5.44, 30.99),$$

exactly the same as with the Binomial data. This follows because of two things. First,  $\sum y_i \sim \text{Bin}(n, \theta)$ , and second  $\sum y_i$  is something called a *sufficient statistic* so that the Bayesian analysis only depends on the data through  $\sum y_i$ , see Section 4.4. If these two things are true, the analysis better agree with the Binomial analysis!

To find the posterior, apply Bayes' Theorem to the densities:

$$\begin{aligned} p(\theta|y) &= \frac{f(y|\theta)p(\theta)}{\int f(y|\theta)p(\theta)d\theta} \\ &= \frac{\theta^{\sum y_i}(1-\theta)^{n-\sum y_i}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}I_{(0,1)}(\theta)}{\int_0^1 \theta^{\sum y_i}(1-\theta)^{n-\sum y_i}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}d\theta} \\ &= \frac{\theta^{\sum y_i+a-1}(1-\theta)^{n-\sum y_i+b-1}I_{(0,1)}(\theta)}{\int_0^1 \theta^{\sum y_i+a-1}(1-\theta)^{n-\sum y_i+b-1}d\theta} \end{aligned}$$

which, using the same trick as with the Binomial, can be seen as the density of a Beta( $\sum y_i + a, n - \sum y_i + b$ ).

Now is a good time to discuss how we arrived at our prior. Suppose that (i) a researcher has estimated that 10% of transportation workers use drugs on the job, and (ii) the researcher is 95% sure that the actual proportion was no larger than 25%. Mathematically, our best guess for  $\theta$  is 0.10 and we have  $\Pr(\theta < 0.25) = 0.95$ .

One way to model the prior assumes that the prior is a member of some parametric family of distributions and uses the two pieces of information to identify an appropriate member of the family. For example, if

$$\theta \sim \text{Beta}(a, b),$$

we might identify the estimate of 10% with either the mode  $(a-1)/(a+b-2)$  [the value at which the pdf achieves its maximum value], or the mean  $a/(a+b)$  of the Beta distribution. We prefer using the mode so we set

$$0.10 = \frac{a-1}{a+b-2}.$$

This allows us to find  $a$  in terms of  $b$ , namely

$$a(b) = \frac{1 + 0.1(b-2)}{0.9}.$$

We then search through  $b$  values until we find a distribution Beta( $a(b), b$ ) for which  $\Pr(\theta < 0.25) = 0.95$ . The Beta(3.44, 22.99) distribution satisfies the constraints. The density is shown in Figure 2.4. Since the prior is based on just two characteristics of a parametric family, once this prior is found, one should go back to the researcher and verify that the entire distribution is a reasonable representation of the researcher's beliefs.

A commonly used Beta prior distribution has  $a = b = 1$ , the uniform distribution with density  $p(\theta) = I_{(0,1)}(\theta)$ . This is a common reference prior and corresponds to a belief that all possible proportions are equally likely, so the expert would be saying that it was equally plausible for the proportion of drug users to be 0.00001 or 0.99999 or 0.5. Such a prior seems inappropriate for drug use since most of us think that  $\theta$  is much more likely to be less than 0.5 than to be above it.

Figure 2.5 gives three Beta densities: the uniform, one concentrated above 0.5 (area to right of 0.5 = 0.875), and another concentrated below 0.5 (area to left of 0.5 = 0.875). The Beta family is rich, encompassing many shapes like the four shown in Figures 2.4 and 2.5. It allows no bimodal shapes but we find that bimodal prior distributions on probabilities are rare. We typically use Beta distributions to model uncertainty about proportions.

**EXERCISE 2.6.** Show that the mode of the Beta( $a, b$ ) distribution is  $(a-1)/(a+b-2)$  when  $a, b > 1$ , zero when  $a < 1$  and one when  $b < 1$ .

Now consider an example using iid normal data with a known variance. The assumption of a known variance is made to simplify the mathematics. It is not usually a realistic assumption when seriously analyzing data.

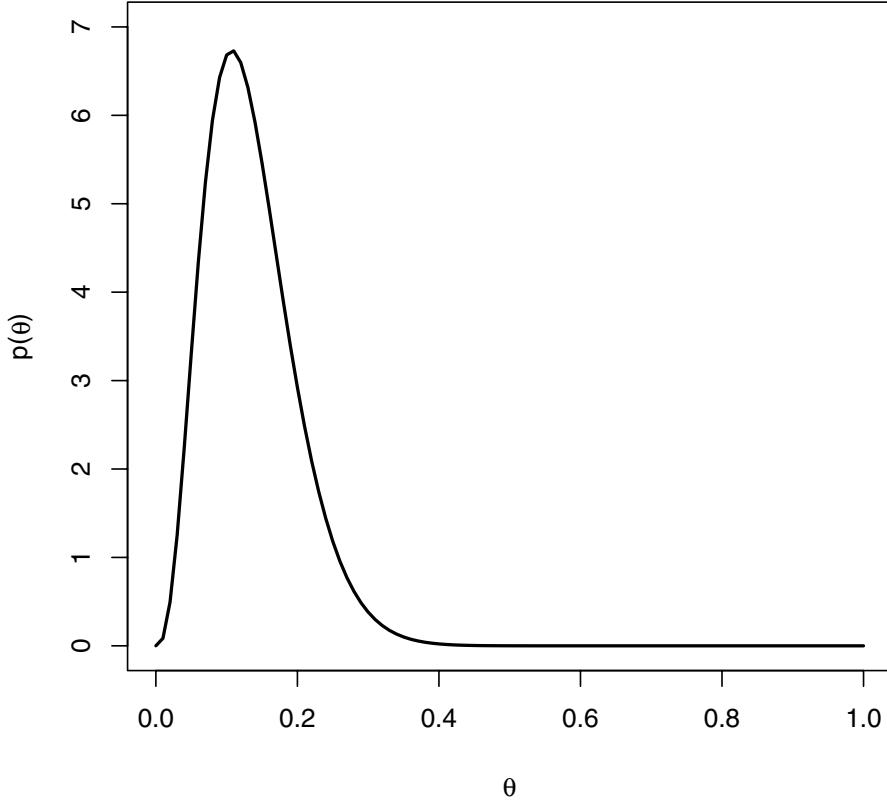


Figure 2.4: Density function of Beta(3.44, 22.99).

**EXAMPLE 2.3.3. Normal Data.** The height of U.S. adult females is assumed to follow a normal distribution with unknown mean  $\theta$  and known variance  $\sigma_*^2$ . A random sample of  $n$  heights is denoted  $y_1, \dots, y_n$ . Here the vector of all the heights  $y = (y_1, \dots, y_n)'$  constitutes the data. The probability model for known variance  $\sigma_*^2$  and  $y_i$ s conditionally independent given the mean  $\theta$  is

$$y_1, \dots, y_n | \theta \stackrel{iid}{\sim} N(\theta, \sigma_*^2).$$

For Bayesians, it is extremely convenient to reexpress the variance as the *precision*

$$\tau_* \equiv 1/\sigma_*^2.$$

With this notation, the density for a single  $y_i$  is

$$f_*(y_i | \theta) = (\sqrt{\tau_*/2\pi}) \exp [-\tau_*(y_i - \theta)^2/2].$$

The density for the random sample is

$$f(y | \theta) = \prod_{i=1}^n f_*(y_i | \theta)$$

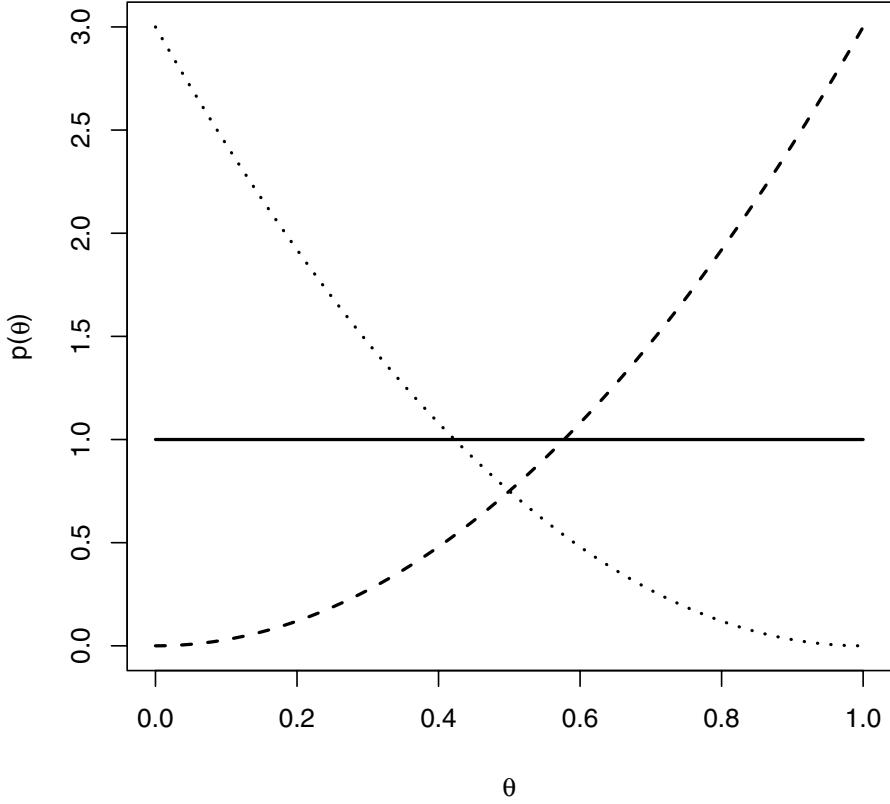


Figure 2.5: Three Beta densities. The flat curve is a Beta(1,1), the curve with mode = 1 is a Beta(3,1), and the curve with mode = 0 is a Beta(1,3).

$$\begin{aligned}
 &= \prod_{i=1}^n (\sqrt{\tau_*/2\pi}) \exp[-\tau_*(y_i - \theta)^2/2] \\
 &= (\sqrt{\tau_*/2\pi})^n \exp \left[ -\tau_* \sum_{i=1}^n (y_i - \theta)^2/2 \right].
 \end{aligned}$$

A simple prior having a density with the same functional form as the data is

$$\theta \sim N(\theta_0, 1/\tau_0).$$

This is both a conjugate prior and a DAP. Write the sample mean as  $\bar{y} = \sum_{i=1}^n y_i/n$ . A considerable amount of algebra is involved in showing that the posterior distribution is

$$\theta|y \sim N \left( \frac{\tau_0}{\tau_0 + n\tau_*} \theta_0 + \frac{n\tau_*}{\tau_0 + n\tau_*} \bar{y}, \frac{1}{\tau_0 + n\tau_*} \right).$$

Note that the posterior mean is a weighted average of the prior mean  $\theta_0$  and the sample mean  $\bar{y}$ . The weights are proportional to the prior precision  $\tau_0$  and to  $n\tau_* = 1/(\sigma_*^2/n)$ , which is the precision of the sample mean. The posterior precision is  $\tau_0 + n\tau_*$ , which is the sum of the prior precision

and the precision of the sample mean. To interpret the prior as a DAP, think of having  $\tau_0/\tau_*$  prior observations from a normal with precision  $\tau_*$  and from these prior observations obtaining a sample mean of  $\theta_0$ , then the posterior precision is  $\tau_*$  times the total sample size  $n + \tau_0/\tau_*$  and the posterior mean is a weighted average of the prior and sample means with weights proportional to the prior and data sample sizes. The prior is conjugate in the sense that both the prior and posterior are normal distributions.

Examples 2.3.1 and 2.3.2 on Binomial and Bernoulli analyses gave the same posterior distribution. A similar result obtains from comparing the full normal data analysis given here with a situation in which one observes only  $\bar{y} \sim N(\theta, 1/n\tau_*)$ .

Most of our interesting data will not be iid, they involve more complicated models. A generic probability model for conditionally independent data, such as regression data, can be written succinctly as

$$y_i|\theta \stackrel{\text{ind}}{\sim} f_i(\cdot|\theta) \quad i = 1, \dots, n$$

in which the observations are independent given  $\theta$ . The right-hand side involves some method of specifying the distribution of  $y_i$ . Here  $f_i(\cdot|\theta)$  denotes the density of  $y_i$ . If  $y_i$  is normally distributed with mean  $\mu_i$  and variance  $\sigma^2$ , rather than specifying the density we might write

$$y_i|\mu_i, \sigma^2 \stackrel{\text{ind}}{\sim} N(\mu_i, \sigma^2) \quad i = 1, \dots, n,$$

where  $\theta = (\mu_1, \dots, \mu_n, \sigma^2)'$ . In either case, the model should correspond to any physical mechanisms that generate the data. Because the  $y_i$ s are conditionally independent, the joint conditional probability density of all the  $y_i$ s is obtained by multiplying the individual conditional densities:

$$f(y|\theta) = \prod_{i=1}^n f_i(y_i|\theta). \quad (3)$$

In some applications the data are not conditionally independent given the parameters, so the sampling density of the data  $f(y|\theta)$  must be constructed differently from (3). In any case, we typically assume some functional form for the sampling density.

## 2.4 Posterior Analysis

In addition to using the sampling density  $f(y|\theta)$ , Bayesians incorporate prior knowledge about  $\theta$  through a density  $p(\theta)$ . The joint density of  $\theta$  and  $y$  is then

$$p(\theta, y) = f(y|\theta)p(\theta).$$

Integrating out  $\theta$ , the marginal density of  $y$  is

$$f(y) = \int f(y|\theta)p(\theta)d\theta.$$

The marginal distribution of the data is sometimes called the *marginal predictive distribution*. By definition, the conditional density of  $\theta$  given  $y$  is

$$p(\theta|y) \equiv \frac{p(y, \theta)}{f(y)}.$$

Bayes' Theorem tells us that

$$p(\theta|y) = \frac{f(y|\theta)p(\theta)}{\int f(y|\theta)p(\theta)d\theta}.$$

The integral in the denominator is  $r$  dimensional. If  $\theta$  has a discrete distribution, the integral is replaced by a sum.

The *posterior density*  $p(\theta|y)$  is a function of  $\theta$ , so the denominator  $f(y)$  is merely a constant. In other words, from probability theory,

$$1 = \int p(\theta|y)d\theta,$$

so the denominator is whatever constant is needed to make  $f(y|\theta)p(\theta)$  integrate to 1. We often write

$$p(\theta|y) \propto f(y|\theta)p(\theta),$$

since the function on the right determines the posterior. In applications, the right-hand side can sometimes be recognized as having the form of a well-known distribution such as a Beta, normal, or gamma density.

*To a Bayesian, the best information one can ever have about  $\theta$  is to know the posterior density  $p(\theta|y)$ .* Nonetheless, it is often convenient to summarize the posterior information. Most summaries involve integration, which we will typically perform by computer simulation as in Chapter 3.

When  $\theta$  is a scalar ( $r = 1$ ) with a continuous distribution, the posterior median, say  $\tilde{\theta} \equiv \tilde{\theta}(y)$ , satisfies

$$\frac{1}{2} = \int_{-\infty}^{\tilde{\theta}(y)} p(\theta|y)d\theta.$$

The posterior mode is the value of  $\theta$ , say  $\theta_M \equiv \theta_M(y)$ , such that  $\max_\theta\{p(\theta|y)\} = p(\theta_M|y)$ , that is,

$$\theta_M \equiv \arg\{\max_\theta[p(\theta|y)]\}.$$

The posterior mean is

$$E(\theta|y) = \int \theta p(\theta|y)d\theta,$$

and the posterior variance is

$$\text{Var}(\theta|y) = \int [\theta - E(\theta|y)]^2 p(\theta|y)d\theta,$$

so the posterior standard deviation is  $\sqrt{\text{Var}(\theta|y)}$ .

Another useful summary is, say, a 95% probability interval  $[a(y), b(y)]$  where  $a(y)$  and  $b(y)$  satisfy

$$0.95 = \int_{a(y)}^{b(y)} p(\theta|y)d\theta.$$

There are many choices for  $a$  and  $b$ , so one might want to find a “best” interval, typically the shortest interval that contains 95% probability. This is called a *highest posterior density (HPD)* interval. For convenience, we typically choose  $a(y)$  and  $b(y)$  so that  $\Pr[\theta < a(y)|y] = 0.025$  and  $\Pr[\theta < b(y)|y] = 0.975$ . Unless otherwise indicated, we call this the 95% *probability interval (PI)*. We can use any percentage in PIs but most often we use 95% or 90%.

In multidimensional problems, the marginal posterior density of, say,  $\theta_1$  and  $\theta_2$  is

$$p(\theta_1, \theta_2|y) = \int \cdots \int p(\theta|y)d\theta_3 \cdots d\theta_r.$$

We might be interested in

$$\Pr(\theta_1 > \theta_2|y) = \int_{-\infty}^{\infty} \int_{\theta_2}^{\infty} p(\theta_1, \theta_2|y)d\theta_1 d\theta_2.$$

We may also be interested in some function of  $\theta$ , say,  $\gamma \equiv g(\theta)$ . The ultimate Bayesian inference about  $\gamma$  is its posterior density,  $p(\gamma|y)$ . More specifically, we might find things like the posterior mean

$$\hat{\gamma} \equiv E(\gamma|y) = \int g(\theta)p(\theta|y)d\theta = \int \gamma p(\gamma|y)d\gamma,$$

variance  $\text{Var}(\gamma|y)$ , median  $\tilde{\gamma}$ , mode  $\gamma_M$ , or the posterior probability of being in a set  $A$ ,

$$\Pr[\gamma \in A|y] = \Pr[g(\theta) \in A|y] = \int_A I_A[g(\theta)]p(\theta|y)d\theta.$$

Consider now prediction of future observations  $\tilde{y} = (\tilde{y}_1, \dots, \tilde{y}_q)'$ . Often these are conditionally independent of  $y$  given  $\theta$ . The same scientific theory that gives us  $f(y|\theta)$  should also provide a density  $f_p(\tilde{y}|\theta)$  for the new observations. The *predictive density* of the future observations given the past observations is

$$f_p(\tilde{y}|y) = \int f_p(\tilde{y}|\theta)p(\theta|y)d\theta.$$

To see this, use the fact that by conditional independence  $f_p(\tilde{y}|y, \theta) = f_p(\tilde{y}|\theta)$  and note that for any (measurable) set  $A$ , by the Law of Total Probability

$$\begin{aligned} \Pr(\tilde{y} \in A|y) &= \int \Pr(\tilde{y} \in A|y, \theta)p(\theta|y)d\theta \\ &= \int \left[ \int_A f_p(\tilde{y}|y, \theta)d\tilde{y} \right] p(\theta|y)d\theta \\ &= \int \left[ \int_A f_p(\tilde{y}|\theta)p(\theta|y)d\tilde{y} \right] d\theta \\ &= \int_A \left[ \int f_p(\tilde{y}|\theta)p(\theta|y)d\theta \right] d\tilde{y} \\ &= \int_A f_p(\tilde{y}|y)d\tilde{y}. \end{aligned}$$

Frequently,  $f_p(\cdot|\theta)$  has a similar functional form to  $f(\cdot|\theta)$  in which case we omit the subscript  $p$ . The predictive distribution is sometimes called the *posterior predictive distribution*.

Suppose that  $\tilde{y}$  is a scalar. We might be interested in, say, the probability that  $\tilde{y} \leq 5$ . It is natural to compute the predictive probability

$$\Pr[\tilde{y} \leq 5|y] = \int_{-\infty}^5 f_p(\tilde{y}|y)d\tilde{y}.$$

We may instead be interested in the parameter  $\Pr[\tilde{y} \leq 5|\theta]$ . As a parameter, we could do many things such as give a 95% probability interval for it. We could also compute the posterior mean as a point estimate,

$$E[\Pr(\tilde{y} \leq 5|\theta)|y].$$

We will see in Section 4.5 that this is the same as the predictive probability.

Fundamentally, it does not really matter whether a scientific theory is correct. (They almost never are.) What matters is whether the scientific theory allows us to make useful predictions about future observations. Better understanding of scientific processes allows us to build better models  $f(y|\theta)$  and  $f_p(\tilde{y}|\theta)$ , which should lead to better predictions. More insightful evaluations of the world (objectivity) should allow more appropriate evaluations of our current state of knowledge  $p(\theta)$ . But ultimately, the test of the quality of our science is how well  $f_p(\tilde{y}|y)$  predicts what we see in the future. Of course, different people will have different evaluations of the state of nature, and perhaps even different scientific models. The closest we ever come to a correct scientific model is when we can make sufficiently accurate predictions by using  $f(y|\theta)$  and  $f_p(\tilde{y}|\theta)$ . We will only ever agree on the state of nature when we have sufficient data so that we essentially agree on  $p(\theta|y)$ , and more usefully  $f(\tilde{y}|y)$ , even though we did not agree on  $p(\theta)$ . See also the discussion on consistency in Section 4.11.

Table 2.1: *Common distributions and densities.*

Distribution	Notation	Density
Bernoulli	Bern( $\theta$ )	$f(y \theta) = \theta^y(1-\theta)^{1-y}; y=0,1$
Binomial	Bin( $n, \theta$ )	$f(y \theta) = \binom{n}{y} \theta^y(1-\theta)^{n-y}; y=0,1,\dots,n$
Beta	Beta( $a,b$ )	$p(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1} I_{(0,1)}(\theta)$
Uniform	$U(a,b)$	$p(\theta) = \frac{I_{(a,b)}(\theta)}{b-a}$
Poisson	Pois( $\theta$ )	$f(y \theta) = \theta^y e^{-\theta}/y!; y=0,1,2,\dots$
Exponential	Exp( $\theta$ )	$f(y \theta) = \theta e^{-\theta y} I_{(0,\infty)}(y)$
Gamma	Gamma( $a,b$ )	$p(\theta) = [b^a/\Gamma(a)]\theta^{a-1}e^{-b\theta} I_{(0,\infty)}(\theta)$
Chi-squared	$\chi_n^2$	Same as Gamma( $n/2, 1/2$ )
Weibull	Weib( $\alpha, \lambda$ )	$f(y \alpha, \lambda) = \lambda \alpha y^{\alpha-1} \exp(-\lambda y^\alpha) I_{(0,\infty)}(y)$
Normal	$N(\mu, 1/\tau)$	$f(y \mu, \tau) = (\sqrt{\tau/2\pi}) \exp[-\tau(y-\mu)^2/2]$
Student's $t$	$t(n, \mu, \sigma)$	$f(y \mu, \sigma) = [1 + (y-\mu)^2/n\sigma^2]^{-(n+1)/2} \times \Gamma[(n+1)/2]/\Gamma(n/2)\sigma\sqrt{n\pi}$
Cauchy	Cauchy( $\theta$ )	Same as $t(1, \theta, 1)$
Dirichlet	Dirch( $a_1, a_2, a_3$ )	$p(\theta) = \Gamma(a_1+a_2+a_3)/\Gamma(a_1)\Gamma(a_2)\Gamma(a_3) \times \theta_1^{a_1-1}\theta_2^{a_2-1}(1-\theta_1-\theta_2)^{a_3-1} \times I_{(0,1)}(\theta_1)I_{(0,1)}(\theta_2)I_{(0,1)}(1-\theta_1-\theta_2)$

## 2.5 Commonly Used Distributions

Prior to the widespread use of computers to approximate posterior distributions, Bayesian statistics could only be performed when the calculus involved in finding the posterior distribution could be performed. This calculus is often relatively simple when, in terms of the parameters  $\theta$ , the prior density  $p(\theta)$  has the same functional form as the sampling density  $f(y|\theta)$  so that, after applying Bayes' Theorem, the posterior has the same functional form as the prior. This was illustrated in Section 3. Such pairs of sampling densities and prior densities are called *conjugate families*.

Table 2.1 gives a list of commonly used distributions and their densities. Densities are denoted either  $f(y|\theta)$  or  $p(\theta)$  depending on whether the distribution is most often used to model the distribution of the data  $y$  given  $\theta$  or the state of nature  $\theta$ . Some distributions, notably the normal and the gamma, are often used for modeling both  $y|\theta$  and  $\theta$ .

There are numerous important relationships between the distributions including

$$\begin{aligned} \text{Bern}(\theta) &= \text{Bin}(1, \theta), & U(0, 1) &= \text{Beta}(1, 1), \\ \text{Exp}(\theta) &= \text{Gamma}(1, \theta), & \text{Exp}(\theta) &= \text{Weib}(1, \theta), \\ N(\mu, 1/\tau) &= t(\infty, \mu, 1/\sqrt{\tau}), & \text{Cauchy}(\theta) &= t(1, \theta, 1), \end{aligned}$$

and

$$\chi_n^2 = \text{Gamma}(n/2, 1/2).$$

We give the Dirichlet distribution in three dimensions but it is easily extended to arbitrary dimensions.

Let  $z$  be a random variable, and for constants  $\mu$  and  $\sigma$ , let  $y = \sigma z + \mu$ . If  $z \sim N(0, 1)$ , then  $y \sim N(\mu, \sigma^2)$  and if  $z \sim t(n) \equiv t(n, 0, 1)$ , then  $y \sim t(n, \mu, \sigma)$ . In both cases,  $\theta = (\mu, \sigma)'$  where  $\theta$  can be redefined with  $\sigma$  replaced by  $\sigma^2$ .

If

$$\log(z) \sim N(\mu, \sigma^2),$$

then  $z$  is said to have a *log-normal distribution* with the same parameters, written

$$z \sim LN(\mu, \sigma^2).$$

Finally, note that if  $\theta$  is a random variable,  $A$  is some set, and  $w = I_A(\theta)$ , then

$$E(w) = \int I_A(\theta) p(\theta) d\theta = \Pr(\theta \in A).$$

**EXERCISE 2.7.** Consider the expression (2.3.3) and define  $f_i(y_i|\theta) \equiv f(y_i|\theta, x_i)$  where  $x_i$  denotes a known “covariate” variable. In particular, let the  $x_i$ s identify two groups,  $x_i = 1$  for  $i = 1, \dots, k$  and  $x_i = 0$  for  $i = k+1, \dots, n$ . With  $\theta = (\theta_1, \theta_2)'$ , define  $\lambda_i = \theta_1^{x_i} \theta_2^{1-x_i}$  so that  $\lambda_i$  equals  $\theta_1$  if  $x_i = 1$  and equals  $\theta_2$  if  $x_i = 0$ . Now identify  $f(y_i|\theta, x_i) \equiv f(y_i|\lambda_i)$  for each of the choices (i)  $f(y_i|\lambda_i) = \lambda_i e^{-\lambda_i y_i}$ , (ii)  $f(y_i|\lambda_i) = \lambda_i^{y_i} e^{-\lambda_i}/y_i!$ , and (iii)  $f(y_i|\lambda_i) = 2\lambda_i y_i e^{-\lambda_i y_i^2}$ . (a) Using Table 2.1, for each of the three choices identify the distribution with density  $f(y_i|\lambda_i)$  when  $x_i = 1$  and also when  $x_i = 0$ . (b) For each choice, simplify the product (2.3.3) so that as a function of  $\theta$  it is proportional to a function that depends only on some combination of  $\sum_{i=1}^k y_i$ ,  $\sum_{i=k+1}^n y_i$ ,  $\prod_{i=1}^k y_i$ , and  $\prod_{i=k+1}^n y_i$ .

Table 2.2 gives means, modes, and variances for the distributions in Table 2.1. Table 2.3 gives commonly used conjugate families for a scalar parameter  $\theta$  in which the posterior distribution is in the same family of distributions as the prior. As illustrated in Section 3, these sampling distributions and conjugate families make for convenient elementary examples when illustrating Bayesian ideas. The first line in Table 2.3 refers to a Beta prior being combined through Bayes’ Theorem with Binomial data  $y$ , resulting in a Beta posterior. The remaining four lines correspond to samples  $y_1, \dots, y_n$  of iid data from (i) Poisson, (ii) exponential, (iii) normal with known variance/precision, and (iv) normal with known mean but unknown variance. As we have seen, the first line could as well have been a sample of Bernoulli random variables. We have already established conjugacy for the Binomial. Conjugacy for the Poisson and exponential are established in Exercise 2.9. The result for a normal with unknown mean but known precision is established in Chapter 4. The normal with known mean and unknown variance is similar to, but easier than, the result established in Chapter 5 for normals with both the mean and the variance unknown.

**EXERCISE 2.8.** Find the median, mode, and mean of the following densities: (i)  $p(\theta) = 2\theta I_{[0,1]}(\theta)$ , (ii)  $f(y|\theta) = \theta e^{-\theta y} I_{[0,\infty)}(y)$ , (iii)  $p(\theta) = I_{[2,4]}(\theta)/2$ . Also find the 90th percentile of each of these distributions, namely, find the value  $c$  such that the area under the density to the left of  $c$  is 0.90 in each case.

**EXERCISE 2.9.** Derive the posterior densities for a Poisson sample with a Gamma prior, and for an exponential sample with a Gamma prior. Show that these densities are as given in Table 2.3.

Table 2.2: *Means, modes, and variances.*

Distribution	Mean	Mode	Variance
Bern( $\theta$ )	$\theta$	0 if $\theta < 0.5$ 1 if $\theta > 0.5$	$\theta(1 - \theta)$
Bin( $n, \theta$ )	$n\theta$	integer closest to $n\theta$	$n\theta(1 - \theta)$
Beta( $a, b$ )	$a/(a + b)$	$(a - 1)/(a + b - 2)$ if $a > 1, b \geq 1$	$\frac{ab}{(a+b)^2(a+b+1)}$
$U(a, b)$	$(a + b)/2$	everything $a$ to $b$	$(b - a)^2/12$
Pois( $\theta$ )	$\theta$	integer closest to $\theta$	$\theta$
Exp( $\theta$ )	$1/\theta$	0	$1/\theta^2$
Gamma( $a, b$ )	$a/b$	$(a - 1)/b$ if $a > 1$	$a/b^2$
$\chi_n^2$	$n$	$n - 2$ if $n > 2$	$2n$
Weib( $\alpha, \lambda$ )	$\Gamma(\frac{\alpha+1}{\alpha})/\lambda \equiv \mu$	$[(\alpha - 1)/\alpha]^{1/\alpha}/\lambda$	$\frac{\Gamma[(\alpha+2)/\alpha]}{\lambda^2} - \mu^2$
$N(\mu, 1/\tau)$	$\mu$	$\mu$	$1/\tau$
$t(n, \mu, \sigma)$	$\mu$ if $n \geq 2$	$\mu$ if $n \geq 3$	$\sigma^2 n / (n - 2)$
Cauchy( $\theta$ )	Undefined	$\theta$	Undefined

Table 2.3: *Some conjugate families.*

$f(y \theta)$	$p(\theta)$	$p(\theta y)$
Bin( $n, \theta$ )	Beta( $a, b$ )	Beta( $a + y, b + n - y$ )
Pois( $\theta$ )	Gamma( $a, b$ )	Gamma ( $\sum y_i + a, n + b$ )
Exp( $\theta$ )	Gamma( $a, b$ )	Gamma( $a + n, b + \sum y_i$ )
$N(\theta, 1/\tau_*)$	$N(\theta_0, 1/\tau_0)$	$N\left(\frac{\tau_0}{\tau_0+n\tau_*}\theta_0 + \frac{n\tau_*}{\tau_0+n\tau_*}\bar{y}_., 1/(\tau_0+n\tau_*)\right)$
$N(\mu_*, 1/\theta)$	Gamma( $a, b$ )	Gamma( $a + n/2, b + n[s^2 + (\bar{y}_. - \mu_*)^2]/2$ ) where $s^2 = n^{-1} \sum (y_i - \bar{y}_.)^2$



---

## Chapter 3

---

# Integration Versus Simulation

---

In Sections 2.3 and 2.4 we illustrated the need to integrate complicated functions when performing Bayesian analysis. Modern computational tools allow us to replace difficult mathematical integrations with simpler computer simulations. This chapter discusses ideas and techniques associated with that replacement. Section 1 introduces ideas of simulation. Section 2 introduces the computer program WinBUGS that we use for most computations. Section 3 discusses a simulation method that is particularly useful in multivariate settings. Section 4 gives some preliminary justifications for simulation methods. Section 5 shows how to perform the binomial example illustrated in this chapter using the computer programming language R.

In statistics, simulation methods are often called *Monte Carlo* methods. A modern computational tool involves simulations based on the probabilistic theory of *Markov chains*. Such simulations are often referred to as *Markov Chain Monte Carlo (MCMC)* methods. Details of simulation techniques are given in Chapter 6.

### 3.1 Introduction

As illustrated in Sections 2.3 and 2.4, virtually all of Bayesian inference involves the calculation of various integrals. Historically, the application of Bayesian methods was limited by one's ability to perform the integrations. Modern Bayesian statistics relies on computer simulations to approximate the values of integrals. The most difficult integrations occur when the dimension  $r$  of the vector  $\theta$  becomes large. Doing high dimensional integrations is almost always difficult, even with simulations.

Our first example uses a very simple model to illustrate integrations that can be performed analytically. The second example illustrates a “simple” problem that is difficult analytically but easy to simulate.

**EXAMPLE 3.1.1. *Simple Analytic Integrations.*** In simple cases, necessary integrations can be performed analytically. Suppose  $\theta$  has prior density  $p(\theta) \propto \theta^2 I_{[0,1]}(\theta)$ . We want to find the proportionality constant  $c$  needed to specify the density, that is, we find  $c$  so that

$$1 = \int_{-\infty}^{\infty} c \theta^2 I_{[0,1]}(\theta) d\theta = \int_0^1 c \theta^2 d\theta.$$

Using calculus,

$$1 = c \frac{\theta^3}{3} \Big|_0^1 = c \frac{1^3}{3} - c \frac{0^3}{3} = c/3,$$

so  $c = 3$ .

The mean of this distribution is

$$E(\theta) = \int_0^1 \theta 3\theta^2 d\theta = \int_0^1 3\theta^3 d\theta = 3 \frac{\theta^4}{4} \Big|_0^1 = \frac{3}{4}.$$

Similar computations give

$$\text{Var}(\theta) = E(\theta^2) - [E(\theta)]^2 = \frac{3}{5} - \left(\frac{3}{4}\right)^2 = 0.0375.$$

The median of this distribution is a number  $\tilde{\theta}$  with

$$0.5 = \int_{-\infty}^{\tilde{\theta}} 3\theta^2 I_{[0,1]}(\theta) d\theta = \int_0^{\tilde{\theta}} 3\theta^2 d\theta.$$

It follows that

$$0.5 = \theta^3 \Big|_0^{\tilde{\theta}} = \tilde{\theta}^3$$

and  $\tilde{\theta} = 0.5^{1/3} = 0.79$ . We can find the 0.025 percentile of  $\theta$  by solving

$$0.025 = \int_0^a 3\theta^2 d\theta$$

for  $a$ . With  $0.025 = \theta^3|_0^a$  or  $a^3 = 0.025$ , solving for  $a$  gives  $a = 0.025^{(1/3)} = 0.292$ . Similarly, we find the 0.975 percentile,  $b = 0.975^{(1/3)} = 0.992$ . We thus have an equal tailed 95% probability interval for  $\theta$ ,  $\Pr(0.292 < \theta < 0.992) = 0.95$ . This probability interval indicates that we are 95% sure  $\theta$  is between about 0.29 and 0.99, which reflects a lot of uncertainty but a strong belief that the parameter is above 0.29.

**EXERCISE 3.1.** Let prior uncertainty about a parameter  $\theta$  be reflected by the density

$$p(\theta) = c e^{-3\theta} I_{(0,\infty)}(\theta).$$

Find the constant  $c$  that makes this integrate to one. Also find  $\Pr(\theta > 2)$  and  $\Pr(\theta > 4 | \theta > 2)$ . Find the median and the expected value. Finally, obtain a 95% probability interval for  $\theta$ .

When analytical solutions are not available, we resort to Monte Carlo simulations.

**EXAMPLE 3.1.2.** *Monte Carlo Integration.* Suppose  $\theta \sim N(0, 1)$ , then

$$\Pr(\theta \geq 1) = \int_1^\infty \frac{1}{\sqrt{2\pi}} e^{-\theta^2/2} d\theta.$$

This integral cannot be evaluated analytically (at least not by us), but we can simulate an answer. Suppose we randomly sample 10,000 observations with  $\theta^k \sim N(0, 1)$ . Consider  $w_k = I_{[1,\infty)}(\theta^k)$ , then the  $w_k$ s are iid Bernoulli random variables with probability of “success,”  $\Pr(\theta \geq 1)$ . To see this formally, note that

$$E(w_k) = \int_{-\infty}^\infty \{I_{[1,\infty)}(\theta)\} \frac{1}{\sqrt{2\pi}} e^{-\theta^2/2} d\theta = \int_1^\infty \frac{1}{\sqrt{2\pi}} e^{-\theta^2/2} d\theta = \Pr(\theta \geq 1).$$

It follows that the sample mean  $\bar{w} = \sum_{k=1}^{10000} w_k / 10,000$  estimates  $\Pr(\theta \geq 1)$ , and since the mean is based on 10,000 samples, it is a very good estimate; one with a standard deviation of

$$\sqrt{\Pr(\theta \geq 1) \{1 - \Pr(\theta \geq 1)\} / 10,000} < 0.005,$$

where the upper bound on the variance is obtained by replacing  $\Pr(\theta \geq 1)$  with 0.5. (It can be shown using calculus that for any probability  $p$ ,  $p(1-p) \leq 1/4$ .) If the standard deviation is too large, simply sample more observations. In fact,  $\Pr(\theta \geq 1)$  is about 0.16, so the actual standard deviation is  $\sqrt{0.16(0.84)/10,000} = 0.0037$ .

Similar to Example 3.1.2, if  $\theta \sim \text{Beta}(a, b)$ , simple probabilities such as

$$\Pr[\theta \geq 0.5] = \int_{0.5}^1 p(\theta) d\theta,$$

although analytically intractable, are easy to simulate.

The approach used in Example 3.1.2 is very general. Suppose we have a posterior density  $p(\theta|y)$  and that we can generate a random sample from the distribution, say  $\theta^1, \theta^2, \dots, \theta^s$ . (We rely on the context to distinguish when, say,  $\theta^2$  is the second component of a sample and when  $\theta^2$  is  $\theta \times \theta$ .) We can approximate the posterior mean using

$$\hat{\theta} \doteq \frac{1}{s} \sum_{k=1}^s \theta^k.$$

Similarly, the posterior variance of a component of the parameter vector, say  $\theta_1$ , can be approximated using the sample variance of the component sample  $\theta_1^1, \theta_1^2, \dots, \theta_1^s$ . Any posterior percentile of  $\theta_1$ , including the median, can be approximated by the corresponding sample percentile. In particular, the approximate 95% probability interval consists of the points between the 2.5 percentile and the 97.5 percentile.

The posterior density can be approximated as a “smoothed” histogram of the sample  $\{\theta^k\}$ . Such smoothing is often done (as in WinBUGS) by using a *kernel smoother*. Associated with the smoother is a window or *bandwidth*. Wider or narrower bandwidths give more or less smoothing. Most software allows the user to pick a bandwidth, but there is typically a default choice. Further details about kernel smoothing would take us too far afield, see Hastie, Tibshirani, and Friedman (2001).

For a function  $\gamma = g(\theta)$ , the posterior mean can be approximated by

$$\hat{\gamma} \doteq \frac{1}{s} \sum_{k=1}^s g(\theta^k).$$

Similarly, for a set  $A$

$$\Pr[\gamma \in A] \doteq \frac{1}{s} \sum_{k=1}^s I_A[g(\theta^k)].$$

By making  $s$  large, we can make these approximations as good as we want. In the last display, adding up the number of random occurrences of a set  $A$  is referred to as the Monte Carlo count. This is not to be confused with the Count of Monte Cristo or his larger brother the Count of Monte Crisco.

Samples from a predictive distribution involve a bit more work. With density  $f_p(\tilde{y}|y) = \int f_p(\tilde{y}|\theta)p(\theta|y)d\theta$ , in addition to sampling  $\theta^k$  from the posterior density  $p(\theta|y)$ , we sample  $\tilde{y}^k$  from the distribution with density  $f_p(\tilde{y}|\theta^k)$ . This is an example of the *method of composition* discussed in Section 3. The sampled  $\tilde{y}^k$  is an observation from the predictive distribution. Do this repeatedly and independently to obtain a Monte Carlo sample  $\{\tilde{y}^k : k = 1, \dots, s\}$ . As usual, the mean of the predictive distribution is numerically approximated by  $\sum_{k=1}^s \tilde{y}^k/s$ , the median of the predictive distribution is approximated by the median of the sample, and an equal tailed 95% predictive probability interval is approximated using the 2.5 and 97.5 sample percentiles.

**EXAMPLE 3.1.3. Two Binomials.** We now expand on Example 2.3.1 by considering two binomials. Assume

$$y_1|\theta_1 \sim \text{Bin}(n_1, \theta_1) \quad \perp \!\!\! \perp \quad y_2|\theta_2 \sim \text{Bin}(n_2, \theta_2).$$

We want to estimate

$$\gamma \equiv \theta_1 - \theta_2,$$

and test  $H_0 : \theta_1 \geq \theta_2$  versus  $H_A : \theta_1 < \theta_2$ , which we rewrite  $H_0 : \gamma \geq 0$  versus  $H_A : \gamma < 0$ .

We need to model the joint prior density  $p(\theta_1, \theta_2)$ . We assume  $\theta_1$  and  $\theta_2$  are independent, so

$$p(\theta_1, \theta_2) = p_1(\theta_1)p_2(\theta_2)$$

with

$$\theta_1 \sim \text{Beta}(a_1, b_1) \quad \perp\!\!\!\perp \quad \theta_2 \sim \text{Beta}(a_2, b_2).$$

Regarding prior independence, the fact that we are comparing these two binomials, in itself, suggests that there is some prior relationship between  $\theta_1$  and  $\theta_2$ . We are not likely to compare the probability that someone will have a heart transplant to the probability that a refrigerator's automatic ice maker will malfunction. We are far more likely to compare the probabilities of ice maker malfunctions for two brands of refrigerators. The key fact to justify the independence assumption is that if we were told the value of  $\theta_1$  (within the plausible range of its prior distribution), we would not want to revise our prior for  $\theta_2$ . We each have some vague idea of how often an ice maker malfunctions and that it does not happen all that often. But if we found out that the first brand of ice maker had a 99% chance of breaking, it would probably change our prior opinion about the probability of the second brand breaking. While the existence of a  $\theta_1$  value (here 99%) that would change our opinions about  $\theta_2$  is technically a violation of independence, this violation is not important. If the hypothetical values for  $\theta_1$  (like 99%) that violate independence are implausible under the current prior, we don't really believe that such values for  $\theta_1$  are possible, so it doesn't matter that they would make us wish to revise the prior on  $\theta_2$ .

The joint posterior density for our independent binomials is

$$\begin{aligned} p(\theta_1, \theta_2 | y_1, y_2) &\propto f(y_1, y_2 | \theta_1, \theta_2)p(\theta_1, \theta_2) \\ &\propto \theta_1^{y_1}(1-\theta_1)^{n_1-y_1}\theta_2^{y_2}(1-\theta_2)^{n_2-y_2}\theta_1^{a_1-1}(1-\theta_1)^{b_1-1}\theta_2^{a_2-1}(1-\theta_2)^{b_2-1} \\ &= \theta_1^{a_1+y_1-1}(1-\theta_1)^{b_1+n_1-y_1-1}\theta_2^{a_2+y_2-1}(1-\theta_2)^{b_2+n_2-y_2-1}. \end{aligned}$$

This can be factored as  $p(\theta_1, \theta_2 | y_1, y_2) = g_1(\theta_1 | y_1)g_2(\theta_2 | y_2)$ , which implies that  $\theta_1$  and  $\theta_2$  are independent given  $y_1, y_2$  and in particular that

$$\theta_1 | y_1 \sim \text{Beta}(a_1 + y_1, b_1 + n_1 - y_1) \quad \perp\!\!\!\perp \quad \theta_2 | y_2 \sim \text{Beta}(a_2 + y_2, b_2 + n_2 - y_2).$$

We want to find the distribution of  $\gamma$ .

The hard way to do this would be analytically by transforming the random vector  $(\theta_1, \theta_2)$  into, say,  $(\gamma, \lambda)$  where  $\lambda \equiv \theta_2$  and using Proposition B.4. Then we could try to find the marginal distribution of  $\gamma$  by integrating out  $\lambda$ . In other words, do a change of variables to find  $p(\gamma, \lambda | y_1, y_2)$  and then find  $p(\gamma | y_1, y_2) = \int p(\gamma, \lambda | y_1, y_2) d\lambda$ . Moreover, for the testing problem we want to find

$$\Pr(\gamma \geq 0 | y_1, y_2) = \int_0^1 p(\gamma | y_1, y_2) d\gamma = \int_0^1 \int_{\theta_2}^1 p(\theta_1, \theta_2 | y_1, y_2) d\theta_1 d\theta_2.$$

There is little possibility of being able to do this analytically.

An easier way to get results is by sampling. Take  $\{(\theta_1^k, \theta_2^k) : k = 1, \dots, s\}$  independently from the posterior distribution. Compute  $\gamma^k = \theta_1^k - \theta_2^k$  and use this to estimate quantities from the posterior distribution of  $\gamma$ . For example, use the sample mean of the  $\gamma^k$ 's to estimate the posterior mean, use the median of the  $\gamma^k$ 's to estimate the median of the posterior distribution, use the 2.5 and 97.5 percentiles of the sample to estimate a 95% probability interval. More generally, we can think of the distribution that takes the value  $\gamma^k$  with probability  $1/s$  as a discrete approximation to the (continuous) posterior distribution. In particular, for the testing problem, define

$$w^k = \begin{cases} 1 & \gamma^k \geq 0 \\ 0 & \text{otherwise} \end{cases},$$

so that

$$\Pr(\theta_1 \geq \theta_2 | y_1, y_2) = \Pr(\gamma \geq 0 | y_1, y_2) \doteq \bar{w} \equiv \sum_{k=1}^s w^k / s.$$

We could also approach the testing problem by examining whether  $\theta_1/\theta_2$  is at least 1. In this approach, let  $\zeta^k = \theta_1^k/\theta_2^k$ , define

$$v^k = \begin{cases} 1 & \zeta^k \geq 1 \\ 0 & \text{otherwise} \end{cases},$$

so that

$$\Pr(\theta_1 \geq \theta_2 | y_1, y_2) = \Pr(\theta_1/\theta_2 \geq 1 | y_1, y_2) \doteq \bar{v} \equiv \sum_{k=1}^s v^k / s.$$

In Section 2 we compare exact Binomial results as developed in Examples 2.3.1 and 3.1.3 to simulation results. To provide another such comparison, Exercises 3.2 and 3.3 derive exact Poisson results that will be compared to simulations in Exercise 3.5.

**EXERCISE 3.2.** Suppose  $n$  cities were sampled and for each city  $i$  the number  $y_i$  of deaths from ALS were recorded for a period of one year. We expect the numbers to be Poisson distributed, but the size of the city is a factor. Let  $M_i$  be the known population for city  $i$  and let

$$y_i | \theta \stackrel{\text{ind}}{\sim} \text{Pois}(\theta M_i), \quad i = 1, \dots, k,$$

where  $\theta > 0$  is an unknown parameter measuring the common death rate for all cities. Given  $\theta$ , the expected number of ALS deaths for city  $i$  is  $\theta M_i$ , so  $\theta$  is expected to be small. Assume that independent scientific information can be obtained about  $\theta$  in the form of a gamma distribution, say  $\text{Gamma}(a, b)$ . Show that this prior and posterior are conjugate in the sense that both have gamma distributions.

$M_i$  is the number of individuals in city  $i$ . It can be written as, say, 100,000 or as 100 thousands. The appropriate units for  $\theta$  depend on how the  $M_i$ s are specified. It can be the rate of events per individual, per 1,000 individuals, or even per million individuals. If we use thousands as units, then a city with 100,000 individuals has  $M_i = 100$  and  $\theta$  is the rate per thousand individuals. If we think 5 out of every 100 thousand individuals is a good guess for  $\theta$ , we want to center our prior  $\text{Gamma}(a, b)$  distribution on 5. On the other hand, when using  $M_i = 100,000$ , the same rate is a number 1,000 times smaller, hence our best guess for  $\theta$  would be  $5/1,000$  and we would center our gamma distribution on that value.

**EXERCISE 3.3.** Extending Exercise 3.2, two cities are allowed different death rates. Let  $y_i \stackrel{\text{ind}}{\sim} \text{Pois}(\theta_i M_i)$ ,  $i = 1, 2$ , where the  $M_i$ s are known constants. Let knowledge about  $\theta_i$  be reflected by independent gamma distributions, namely  $\theta_i \sim \text{Gamma}(a_i, b_i)$ . Derive the joint posterior for  $(\theta_1, \theta_2)$ . Characterize the joint distribution as we did for sampling two independent binomials. Think of  $\theta_i$  as the rate of events per 100 thousand people in city  $i$ . For independent priors  $\theta_i \sim \text{Gamma}(1, 0.1)$ , give the exact joint posterior with  $y_1 = 500, y_2 = 800$  in cities with populations of 100 thousand and 200 thousand, respectively.

### 3.2 WinBUGS I: Getting Started

WinBUGS is a menu driven Windows program that generates samples from the posterior distribution. The samples are not necessarily random samples in the sense of being iid from the posterior. They are identically distributed from the posterior but they are usually dependent samples. Typically,

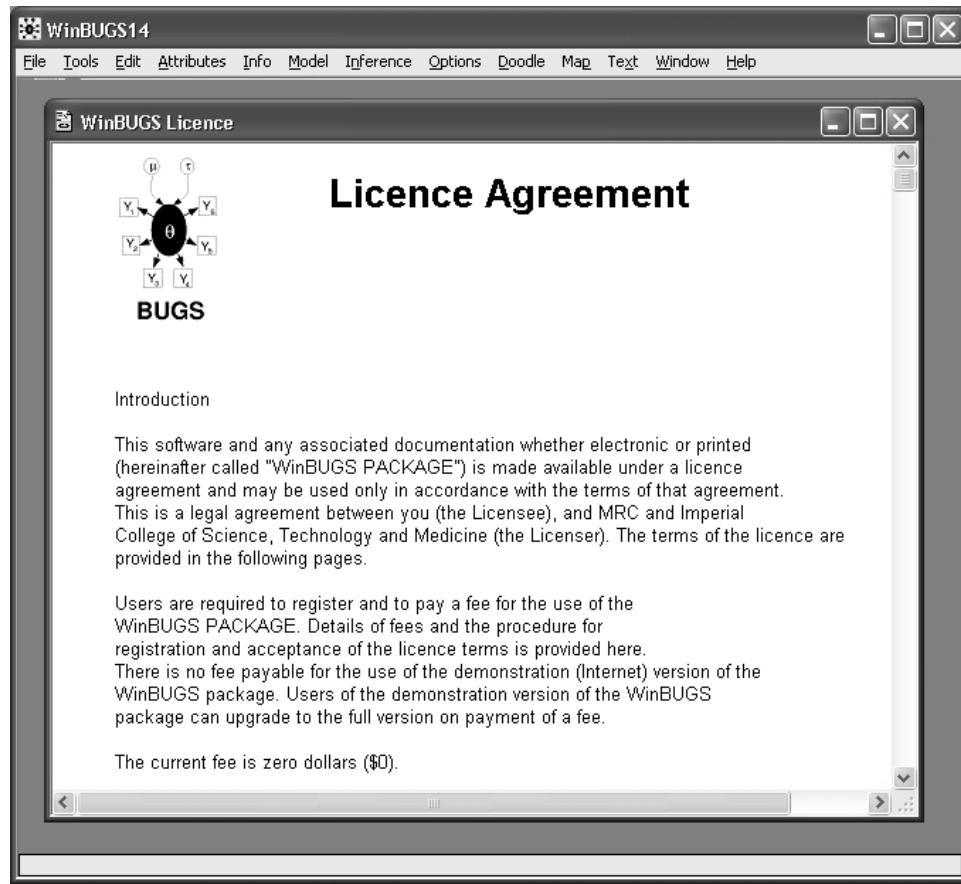


Figure 3.1: *WinBUGS*.

they can be *thinned* to give approximate random samples. In any case, as discussed in Section 4, this lack of independence will not affect most applications. Although menu driven, WinBUGS is not self explanatory.

WinBUGS uses the probabilistic idea of a Markov chain to generate samples. We will need to check whether the Markov chain is behaving so that we can be confident of getting a valid sample from the posterior. We begin with the basics and gradually introduce more sophisticated computing ideas as we go along. The following is a tutorial and the reader will maximize their benefit from it if they literally follow the steps while they are on their computer.

I. Go to the website

<http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml>

and download WinBUGS — assuming you are in a category of people who can do so legally. Note that there is a link on the WinBUGS webpage to a Flash “movie” illustrating the steps involved in getting a WinBUGS program to run. The movie uses separate windows for the WinBUGS code, data, and initial values. We place everything in one window in the example that follows.

II. Having installed WinBUGS, double click on the WinBUGS icon or in some other way start WinBUGS running.

- III. After reading every word of the license agreement, kill the window it came up in (see Figure 3.1). What commitments have you made regarding your firstborn?
- IV. You can either proceed to our binomial example or open the “Help” menu and then open the “User Manual” option. Explore the contents of the User Manual. There is a tutorial included that guides the user step-by-step through the process of running a WinBUGS program and obtaining inferences. There are also two volumes of WinBUGS examples including code, data, initial values, and sample output. Often, these examples can be tapped as a starting point for fitting a similar model to your own data. Kill the “User Manual” window.
- V. Mentally send your thanks to David Spiegelhalter, Andrew Thomas, Nicky Best, and everyone else responsible for the existence and distribution of WinBUGS. For additional information on WinBUGS see Lunn et al. (2000) and Lunn et al. (2009).

You are now ready to enter the world of Markov Chain Monte Carlo (MCMC) approximations to Bayesian posterior distributions.

We step through a simple WinBUGS example. The last time one of us taught linear model theory, 7 of the 10 students were taller than 67 inches. We consider 7 to be the observed outcome of a binomial random variable with 10 trials. Our parameter of interest is  $\theta$ , the probability that a linear models student will be taller than 67 inches. We assume a uniform prior on  $\theta$ , that is, a Beta(1,1), which is a standard reference prior. As discussed in Example 2.3.1, for this problem the mathematics can be worked out. The posterior distribution is Beta(8,4). The mean of the posterior distribution is  $8/12 = 0.667$ , the variance is  $(8/12)(4/12)/13 = (0.1307)^2$ . Let’s see how WinBUGS performs on these data.

In analyzing the data, the WinBUGS14 window will generate numerous sub-windows, see Figure 3.2.

- I. Go to the “File” menu and click on “New.” This opens a window entitled “untitled1.” (In Figure 3.2, the corresponding window is “SimpleBinomial.”)
- II. In this new window, specify the model  $y|\theta \sim \text{Bin}(10, \theta)$ ,  $\theta \sim \text{Beta}(1, 1)$ , the data  $y = 7$ , and a starting value for the computations. To do this, type the WinBUGS commands:

```
model;
{ y ~ dbin(theta,10)
  theta ~ dbeta(1,1) }
list(y=7)
list(theta=.5)
```

The WinBUGS code is in the braces following the `model` statement. Either `model;` followed by the model specification in braces, or simply `model{...}` works to specify the probability model.

In WinBUGS the specification for a binomial has the parameter first, then the number of trials, which reverses the order from standard notation for specifying binomials. The command “`list(theta=.5)`” sets an initial value for the computation. Now would be a good time to save the WinBUGS code in “untitled1” as an “odc” file in case your computer crashes. Use the WinBUGS14 “File” menu. We called our file `SimpleBinomial.odc` so that `SimpleBinomial` becomes the name of this window.

- III. Back in the WinBUGS14 window, open the menu “Model” and click on “Specification...” This opens a new “Specification Tool” window.

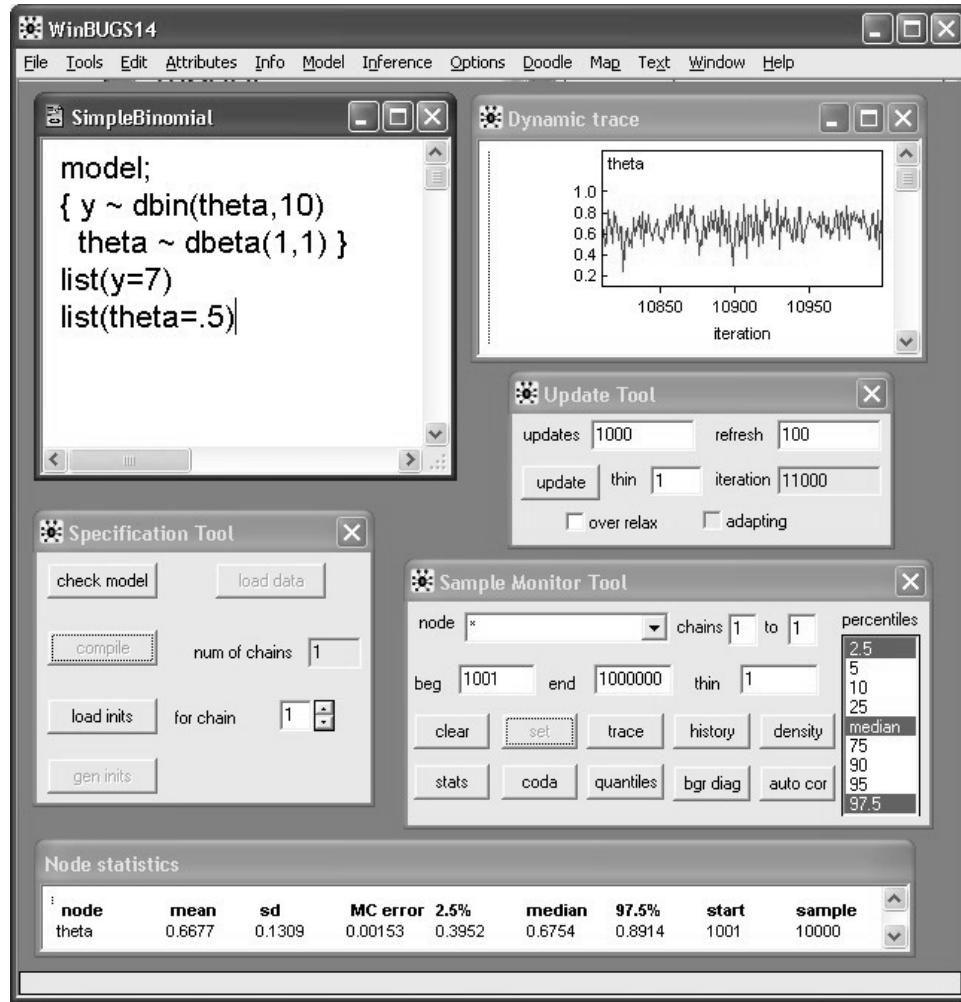
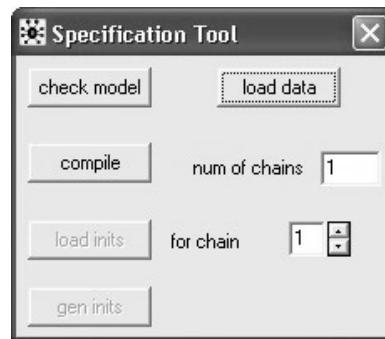
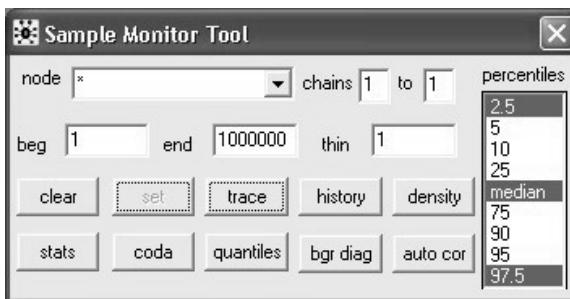


Figure 3.2: *WinBUGS after fitting binomial model.*



We will need to go back and forth between the “Specification Tool” and “SimpleBinomial” windows. In the bottom left of the WinBUGS14 window, it will tell you what is going on, that is, whether things are working properly and, if not, what WinBUGS thinks is going wrong. WinBUGS uses a vertical line as a cursor, and if something is wrong, WinBUGS puts a faded cursor where it thinks the problem exists. (We frequently manage to fool WinBUGS with our errors.)

- IV. Double click in the middle of the word “model” in “SimpleBinomial” then click on **check model** in the “Specification Tool” window.
- V. Double click in the middle of the word “list” in the line with  $y = 7$  in “SimpleBinomial” then click on **load data** in “Specification Tool.”
- VI. Click on **compile** in “Specification Tool.”
- VII. Double click in the middle of the word “list” in the line with  $\text{theta}=.5$  in “SimpleBinomial” then click on **load inits** in “Specification Tool.” This gives WinBUGS a starting value for its iterative procedure.
- VIII. Normally you could kill the “Specification Tool” window at this point but we left it open in Figure 3.2.
- IX. Back in the WinBUGS14 window, open the “Inference” menu and choose “Samples...” which opens the “Sample Monitor Tool” window.
- X. In the box by “node” enter “theta” and click on the **set** button. This box tells WinBUGS what quantities (parameters) you want to evaluate in your analysis. To tell WinBUGS you are done entering nodes, put an asterisk in the box. Our window looked like



- XI. Go back to the “Model” menu and choose “Update...” to open the “Update Tool” window.



- XII. Go to “Update Tool.” In the box next to “updates” change the number to 11000. Click the **update** button in the “Update Tool.” When the number in the box next to “iteration” reaches 11000, go back to the “Sample Monitor Tool.”
- XIII. A “burn in” value can be specified in the “beg” box of the “Sample Monitor Tool.” Type “1001” in this box. We will discuss the notion of burn in later, but roughly, we are throwing out the first 1000 iterates to eliminate any effect of our starting value  $\text{theta}=.5$  on posterior inferences. (Specifying a burn-in period is standard, although in this simple binomial example it is not necessary.)
- XIV. In the “Sample Monitor Tool” click on **stats** and **density**.

When we followed these instructions we got the following results from the “Node statistics” window:

node	mean	sd	MC error	2.5%	median	97.5%
theta	0.6677	0.1309	0.00153	0.3952	0.6754	0.8914

along with numbers for “start” and “sample.” Start gives the iteration number of the first observation in the Markov chain that is used to approximate the posterior distribution, while sample gives the number of observations from the chain that are used in the approximation. The other numbers in the table are WinBUGS’s approximations to the posterior mean and posterior standard deviation, an approximation error, and approximations to the posterior 2.5 percentile, median, and 97.5 percentile. For comparison, recall that we know the true posterior mean is  $2/3$  and the true posterior standard deviation is 0.1307. The “Kernel density” window provides an approximation to the Beta(8,4) posterior density.

It is crucial to evaluate how well the MCMC procedure works. It may be difficult to ascertain that it is working well, but one should at least know when it is obviously working badly. We now introduce WinBUGS tools for this purpose. The further discussion of these tools in Subsection 6.3.5 should also be examined, even if readers do not make it through the rest of Chapter 6. These tools can and should be applied to all applications including our earlier binomial example.

Kill all the windows except WinBUGS14. Go to the “File” menu and select “New.” This opens a new window “untitled2.” Type in the following WinBUGS code.

```
model{
  m1 ~ dnorm(0,1)
  m2 ~ dnorm(0,1)
  m3 ~ dnorm(0,1)
}
```

This is merely code to have WinBUGS give us three independent samples from a  $N(0,1)$  distribution. As before, go to the “Specification Tool” window from the “Model” menu and check the model and compile. There are no data, so there is no need to load data. Instead of clicking [load inits], click [gen inits] and then kill the “Specification Tool” window — assuming everything went right. Now, as before, go to “Sample Monitor Tool” and set the nodes `m1`, `m2`, and `m3`. Go to the “Update Tool” and click the [update] button, then go back to the “Sample Monitor Tool” but this time click the [history] button. This generates a new window that has three plots in it. These are examples of the way that histories *should* look.

Often, the beginning of a simulation process does not look like these plots, but if you run enough iterations, the end of the process will look like these. In such cases, you should cut off the beginning of the process and only use the end part that looks good. Cutting off the beginning is referred to as using a “burn in” period. If in looking at the history you think the process settles down after, say, 500 iterations, you can eliminate the first 500 by specifying 501 in the “beg” box of the “Sample Monitor Tool.” (“beg” for begin.) In some cases, it can take many iterations for the process to settle down. Having changed the number in the “beg” box, click the [stats] button and note the values of “start” and “sample” on the extreme right.

In the binomial example, rather than jumping directly to a sample of size 10,000, it is instructive to see how the Markov chain results change as the sample size increases. It is also possible to perform the visual evaluation of the Markov chain in real time. Kill all the windows except WinBUGS14. From the file menu, open SimpleBinomial and repeat steps 3 through 9 of the binomial example. If you have not earlier killed the SimpleBinomial window, just begin again at step 3.

- I. The path of the Markov chain can be monitored dynamically by clicking the [trace] button in the “Sample Monitor Tool.”
- II. In the “Model” menu choose “Update...” to open the “Update Tool” window.
- III. Click the [update] button in the “Update Tool.” This gives a sample of 1,000 observations. The trace plot from step 1 dynamically shows the actual MCMC iterates being generated. In the “Sample Monitor Tool,” click [stats], [density], and [history].

- IV. Again click the **update** button in the “Update Tool.” In addition to the trace plot, in the “Sample Monitor Tool,” again click **stats**, **density**, and **history** to see how they have changed.
- V. Play around! Go crazy! Click **update** several times. Be careful, this can be quite hypnotic. Update the Markov chain until 11000 iterates are generated, i.e. until 11,000 is in the “iteration” box. You can update more or less than 1,000 at a time by typing a different value in the “updates” box.

In the “Sample Monitor Tool,” there is an **auto cor** button. This gives the estimated autocorrelation functions for each node. The autocorrelation function is a standard tool in time series, see Shumway and Stoffer (2006). It measures lack of independence (correlation) over time differences in a sequence of observations. The correlation at time difference (lag) 0 will always be 1. If the rest of the correlations are very near 0, the sequence is acting like a random sample. Suppose that only the first 5 autocorrelations (not including lag 0) are substantial, then we may obtain an approximate random sample by looking at every sixth number in the sequence. To do this, start all over but in the “Update Tool” put a 6 in the “thin” box. After thinning the sequence, it is wise to recompute autocorrelations to make sure that all autocorrelations are near 0 for the thinned sequence.

In the “Update Tool,” you can see the “iteration” box changing when you click the **update** button. The rate at which the number in the iteration box changes is determined by the “refresh” box. As it stands, every time another 100 iterates have been sampled the iteration box changes. You can change this value to 1,000 or 10,000, etc. Typically, it takes far longer to write out the number of iterations to the monitor than it takes to actually do the iterations. When a complicated model causes iterations to be slow, increasing the “refresh” number often lets you get very large samples in very little time. You can get a sense of how long it will take to get your total number of iterations by seeing how fast this number changes while you are running your program.

Examples 3.1.2 and 3.1.3 involved the use of random variables defined by indicator functions. WinBUGS includes a *step function* **step(v)** that is useful for such computations:

$$\text{step}(v) = \begin{cases} 1 & v \geq 0 \\ 0 & v < 0 \end{cases}.$$

For any number  $a$

$$\text{step}(v - a) = \begin{cases} 1 & v \geq a \\ 0 & v < a \end{cases}.$$

For example, to execute Example 3.1.2, use a model that assigns the result of the step function to  $w$

```
model{
  theta ~ dnorm(0,1)
  w <- step(theta-1)
}
```

and use  $w$  as a node. Theoretically,  $\Pr(\theta \geq 1) = 0.1587$ . Somewhat disturbingly, a WinBUGS run of 600,000 samples settled down around 0.1575 for the average value of  $w$ . But as they say, this is still good enough for government work. (Minitab simulations executed in a similar manner for 10,000 and 30,000 samples did worse than comparable stages of the WinBUGS simulation. Some quite extensive simulations by a student suggest that this may be a problem with random number generators. In particular,  $\Pr(\theta < -1)$  works much better.)

The step function must be applied carefully to discrete random variables since

$$\text{step}(v - a) = I_{[a, \infty)}(v) \neq I_{(a, \infty)}(v).$$

If  $v$  is a continuous random variable, all is well

$$\Pr[v > a] = E[I_{(a, \infty)}(v)] = E[I_{[a, \infty)}(v)] = \Pr[v \geq a].$$

However, when  $v$  is discrete taking on integer values, an adjustment must be made to the step function

$$\Pr[v > a] = \mathbb{E}[I_{(a,\infty)}(v)] = \mathbb{E}[I_{[a+1,\infty)}(v)] = \Pr[v \geq a+1].$$

**EXERCISE 3.4.** Perform Example 3.1.3 in WinBUGS with  $y_1 \sim \text{Bin}(80, \theta_1)$ ,  $y_2 \sim \text{Bin}(100, \theta_2)$ ,  $\theta_1 \sim \text{Beta}(1, 1)$ ,  $\theta_2 \sim \text{Beta}(2, 1)$  with observations  $y_1 = 32$  and  $y_2 = 35$ . Put each term in the model on a separate line. There should still be only two list statements with entries separated by commas. See Exercises 3.6 and 3.7 for WinBUGS syntax.

**EXERCISE 3.5.** Perform a data analysis for the model in Exercise 3.3 using the data  $y_1 = 500, y_2 = 800, M_1 = 100, M_2 = 200$ , and using independent Gamma( $1, 0.01$ ) priors for the  $\theta_i$ s. Make WinBUGS based inferences for all parameters and functions of parameters discussed there using a Monte Carlo sample size of 10,000 and a burn-in of 1,000. This may involve an excursion into the “Help” menu to find the syntax for Poisson and gamma distributions. Compare the posterior means for  $\theta_1$  and  $\theta_2$  based on the WinBUGS output to the exact values from the Gamma posteriors that you obtained in Exercise 3.3.

### 3.3 Method of Composition

We now examine simulating from joint distributions that are specified conditionally. We illustrate the method using a combination of normal and gamma distributions. The method is particular useful in sampling predictive distributions.

The technique, known as the *method of composition*, involves simulating from a series of conditional distributions. Consider a joint distribution defined by an observation

$$y|\mu, \tau \sim N(\mu, 1/\tau)$$

and a conditionally specified prior

$$\mu|\tau \sim N(\mu_0, 1/\tau) \quad \tau \sim \text{Gamma}(a, b).$$

The joint density is  $f(y|\mu, \tau)p(\mu|\tau)p(\tau)$ . To get a random sample from the joint distribution, take a random sample

$$\tau^k \sim \text{Gamma}(a, b), \quad k = 1, \dots, s$$

then, conditional on the  $\tau^k$ s, randomly sample

$$\mu^k \sim N(\mu_0, 1/\tau^k),$$

and finally sample

$$y^k \sim N(\mu^k, 1/\tau^k).$$

For  $k = 1, \dots, s$ , the  $(\tau^k, \mu^k, y^k)$ s form a random sample from the joint distribution. For example, one could smooth a histogram of the  $y^k$ s to get a picture of the marginal density of  $y$  and one can directly use the Monte Carlo sample to approximate means or variances.

In WinBUGS, the method of composition is easy to perform. Suppose in our example that  $\tau \sim \text{Gamma}(5, 10)$ , and  $\mu|\tau \sim N(3, 1/\tau)$ . Specify the model as

```
model{
  tau ~ dgamma(5,10)
  mu ~ dnorm(3,tau)
  y ~ dnorm(mu,tau)
}
```

Note that WinBUGS parameterizes normal distributions in terms of the mean and precision rather than the variance. Any of `tau`, `mu`, and `y` may be nodes of interest in the simulations.

Predictive distributions are typically specified via

$$f_p(\tilde{y}|y) = \int f_p(\tilde{y}|\theta)p(\theta|y)d\theta.$$

Thus, to sample from the predictive distribution, sample  $\theta^1, \dots, \theta^s$  from the posterior  $p(\theta|y)$  and for each  $\theta^k$  sample a  $\tilde{y}^k$  from  $f_p(\tilde{y}|\theta)$ . Marginally the  $\tilde{y}^k$ 's are sampled from the predictive distribution. Together the  $(\theta^k, \tilde{y}^k)$ 's are sampled from the joint posterior and predictive distribution  $p(\theta, \tilde{y}|y)$ .

**EXERCISE 3.6.** For Example 2.3.1, use WinBUGS to obtain the predictive discrete density of a future binomial observation  $\tilde{y}$  and give a numerical approximation to  $\Pr(\tilde{y} \geq 20|y)$ . In particular, use  $n = m = 100$ ,  $y = 10$ ,  $\theta \sim \text{Beta}(1, 1)$ , and the following WinBUGS code:

```
model{
  y ~ dbin(theta, n) # Model the data
  ytilde ~ dbin(theta, m) # Prediction of future binomial
  theta ~ dbeta(a, b) # The prior
  prob <- step(ytilde - 20) # Pred prob that ytilde >= 20
}
list(n=100, m=100, y=10, a=1, b=1) # The data
list(theta=0.5, ytilde=10) # Starting/initial values
```

Run this code using a simulation sample size  $s = 5,000$  and a burn-in of 100.

**EXERCISE 3.7.** Consider a two-binomial problem with independent Beta priors as in Example 3.1.3. The following code is designed to obtain the joint posterior for  $(\theta_1, \theta_2)$ , to numerically approximate the posterior probability that  $\theta_1 > \theta_2$ , and to obtain a numerical approximation to the joint predictive density for a pair of future binomials, one from each population. We assume that the future values  $(\tilde{y}_1, \tilde{y}_2)$  are independent of the data  $(y_1, y_2)$  given the parameters  $(\theta_1, \theta_2)$ . For fun, we consider numerically approximating the predictive probability that  $\tilde{y}_1 > \tilde{y}_2 + 10$ . Here's the code:

```
model{
  y1 ~ dbin(theta1, n1) # Modeling the data as
  y2 ~ dbin(theta2, n2) # independent binomials
  theta1 ~ dbeta(1, 1) # Specifying the priors
  theta2 ~ dbeta(1, 1)
  prob1 <- step(theta1-theta2) # Pr(theta1 >= theta2|data)
  y1tilde ~ dbin(theta1, m1) # Get predictive densities
  y2tilde ~ dbin(theta2, m2)
  prob2 <- step(y1tilde - y2tilde - 11)
}
list(y1=25, y2=10, n1=100, n2=100, m1=100, m2=100) # The data
list(theta1=.5,theta2=.5,y1tilde=20,y2tilde=20) #Starting values
```

Run the code with different Monte Carlo sample sizes and burn-in sizes. Modify the code to get other inferences that may be of interest. In particular, obtain the posterior density and a 95% probability interval for  $\gamma = \theta_1 - \theta_2$ .

### 3.4 Monte Carlo Integration\*

We have discussed Monte Carlo integration without offering any justification that it works. Chapter 6 discusses Monte Carlo methods more carefully. Here we introduce the ideas that justify simulation methods.

Think of a study to estimate the average height of all adults living in Kalamazoo, Michigan. Call this unknown mean  $\mu$ . A standard approach would be to take a random sample of size  $n$  from the population, say  $y_1, \dots, y_n$ , and estimate  $\mu$  by the sample mean  $\bar{y}$ . Before we actually go out and get the sample,  $\bar{y}$  is random (because the sampling is random) and  $\bar{y}$  has a distribution. Intuitively, larger  $n$  gives a better  $\bar{y}$  for estimating  $\mu$ . This is because for larger  $n$  more information (i.e., more heights) goes into the estimate. So *typically*, larger  $n$  yields  $\bar{y}$ s closer to  $\mu$ . In the extreme case when  $n$  is as large as possible and the entire population is sampled,  $\bar{y} = \mu$ .

Restating this fundamental idea in words: Averages derived from larger and larger random samples get arbitrarily close to the true (population) average with high probability. A theorem from probability called the *Law of Large Numbers (LLN)* makes this idea concrete.

**Proposition 3.1. (LLN)** Let  $\theta^1, \theta^2, \dots$  be iid with density  $p(\theta)$ , then

$$\overline{g(\theta)}_s \equiv \frac{1}{s} \sum_{k=1}^s g(\theta^k) \xrightarrow{P} E[g(\theta)] \equiv \int g(\theta)p(\theta)d\theta.$$

The symbol  $\xrightarrow{P}$  is read “converges in probability.” Simply stated, this means that as larger samples are taken, the probability approaches zero that the sample mean on the left side is more than any given (arbitrarily small) distance away from  $E[g(\theta)]$ . All our results are stated for sampling from the prior distribution of  $\theta$  but they apply immediately to the posterior as well.

The LLN is more flexible than you might think. Not only does it ensure that

$$\bar{\theta}_s \xrightarrow{P} E(\theta)$$

but squaring each  $\theta^k$ , we get

$$\frac{1}{s} \sum_{k=1}^s [\theta^k]^2 \xrightarrow{P} E[\theta^2].$$

When  $\theta$  is a scalar, indirectly we get the sample variance converging to the population variance because  $s/(s-1) \rightarrow 1$  and

$$\frac{1}{s} \sum_{k=1}^s [\theta^k - \bar{\theta}_s]^2 = \frac{1}{s} \sum_{k=1}^s [\theta^k]^2 - [\bar{\theta}_s]^2 \xrightarrow{P} E[\theta^2] - [E(\theta)]^2 = \text{Var}(\theta).$$

Similarly for the standard deviations,

$$\sqrt{\frac{1}{s} \sum_{k=1}^s [\theta^k - \bar{\theta}_s]^2} \xrightarrow{P} sd(\theta).$$

The LLN also gives, for example,

$$\frac{1}{s} \sum_{k=1}^s I_{[1, \infty)}(\theta^k) \xrightarrow{P} E[I_{[1, \infty)}(\theta)] = \Pr[\theta \geq 1].$$

Heuristically, the fact that  $s^{-1} \sum_{k=1}^s I_{(-\infty, a]}(\theta^k) \xrightarrow{P} \Pr(\theta \leq a)$  assures that the empirical quantiles of  $\{\theta^k\}$  estimate the posterior quantiles. Although the LLN does not guarantee this, because we cannot write this statement in terms of an integral, Cramér (1946, Sec. 28.5) shows that as  $s \rightarrow \infty$ ,

$$\text{med}\{\theta^k : k = 1, \dots, s\} \xrightarrow{P} \text{med}(\theta) \equiv \tilde{\theta},$$

where the left-hand side refers to the sample median of the Monte Carlo sample and the right-hand side refers to the number that has 0.5 area to the left and 0.5 area to the right under the distribution

of  $\theta$ . Similarly, the  $\alpha/2$  and  $1 - \alpha/2$  percentiles in the Monte Carlo sample converge in probability to the corresponding percentiles of the distribution for  $\theta$ . So if we let  $\hat{\theta}_{0.025}$  and  $\hat{\theta}_{0.975}$  be the sample percentiles and let  $\theta_{0.025}$  and  $\theta_{0.975}$  be the true percentiles of the posterior, then we have

$$0.95 = \Pr(\theta_{0.025} < \theta < \theta_{0.975} | y) = \int_{\theta_{0.025}}^{\theta_{0.975}} p(\theta | y) d\theta \doteq \int_{\hat{\theta}_{0.025}}^{\hat{\theta}_{0.975}} p(\theta | y) d\theta.$$

We thus have  $(\hat{\theta}_{0.025}, \hat{\theta}_{0.975})$  as a numerical approximation to an exact 95% posterior probability interval for  $\theta$ .

Unfortunately Markov chain methods do not give random samples. We usually get a sample  $\theta^1, \dots, \theta^s$  that are identically distributed but *not* necessarily independent. Fortunately, an alternative version of the LLN called the Ergodic Theorem exists for identically distributed but dependent sequences.

The LLN and Ergodic Theorem give us a basis for believing that our simulations provide good answers for large Monte Carlo samples, but how large do the samples need to be? Another theorem, called the *Central Limit Theorem (CLT)*, provides an answer for iid samples. The CLT says that sample means computed from random samples closely follow a normal distribution for large sample sizes.

**Proposition 3.2 (CLT)** Let  $\theta^1, \theta^2, \dots$  be iid with density  $p(\theta)$ , then for large  $s$  we get the approximation

$$\overline{g(\theta)}_s \equiv \frac{1}{s} \sum_{k=1}^s g(\theta^k) \stackrel{\sim}{\sim} N(E[g(\theta)], \text{Var}[g(\theta)]/s).$$

Read the symbol  $\stackrel{\sim}{\sim}$  as “is approximately distributed as.” The CLT allows us to compute a standard error (or Monte Carlo error) for the estimate  $\overline{g(\theta)}_s$ , namely the estimated standard deviation of  $g(\theta)$  divided by  $\sqrt{s}$ . This works out to

$$se(\overline{g(\theta)}_s) = \sqrt{s^{-2} \sum_{k=1}^s [g(\theta^k) - \overline{g(\theta)}_s]^2}.$$

As our samples are not typically iid, WinBUGS uses an alternative method to compute the Monte Carlo error, see Roberts (1996; p. 50).

### 3.5 Posterior Computations in R

In Section 2 we used WinBUGS for posterior computations in a binomial problem with a uniform, i.e., Beta(1, 1) prior on the success probability  $\theta$ . The example had  $y = 7$  successes out of  $n = 10$  independent trials. The posterior distribution is  $\theta | y \sim \text{Beta}(a + y, b + n - y)$ , where  $a = b = 1$  are the hyperparameter values of the beta prior. The same calculations can be performed using the freely available R statistical software package. Appendix C contains an overview of R for data analysis.

The following R code provides an alternative to fitting the model in WinBUGS. We generated a Monte Carlo sample of 10,000 iterates directly from the known posterior. As mentioned earlier, in Figure 3.2 a burn-in of 1,000 iterates was used but was not needed. That is because, in this simple example, WinBUGS also samples directly from the known posterior. The code specifically asks for the posterior mean, standard deviation, median, and 95% PI.

```
y=7
n=10
a=1
b=1
posta=a+y
postb=b+n-y
```

```

MC=10000
posterioriterates=rbeta(MC,posta,postb)
postmean=mean(posterioriterates)
postsd=sd(posterioriterates)
postmed=median(posterioriterates)
postPI=quantile(posterioriterates,c(0.025,0.975))
print(c(postmean,postsd,postmed,postPI))

```

This code generates the output

```
0.6644137 0.1311739 0.6724793 0.3931120 0.8923948
```

The output is in close agreement with the WinBUGS results displayed at the bottom of Figure 3.2. Note that, in general, the R function `rbeta(n, a, b)` returns a random sample of  $n$  values from the Beta( $a, b$ ) distribution.

We also present a second approach to obtaining the posterior summaries, one that does not involve Monte Carlo simulation but instead uses analytical calculations for the known posterior beta distribution.

```

postmean1=posta/(posta+postb)
postsd1=sqrt(posta*postb / ((posta+postb)^2*(posta+postb+1)))
postmed1=qbeta(0.50, posta,postb)
postPI1=qbeta(c(0.025,0.975),posta,postb)
print(c(postmean1,postsd1,postmed1,postPI1))

```

This code generates the output

```
0.6666667 0.1307441 0.6761955 0.3902574 0.8907366
```

In general, the R function `qbeta(p, a, b)` returns the  $p$ th quantile of the Beta( $a, b$ ) distribution. In our application, the code specified a 2-vector of  $p$  values and returned a 2-vector of quantiles. The posterior mean and standard deviation were computed using formulas in Table 2.2. Again, the numbers are close to those reported by WinBUGS.

---

## Chapter 4

---

# Fundamental Ideas II

---

The topics in this chapter are fundamental to the theory and application of Bayesian statistics. A well educated Bayesian should not be ignorant of them. The topics covered are statistical testing, exchangeability, likelihood functions, sufficient statistics, predictivism, Bayes factors and other model selection criteria, large sample normal approximations, consistency, hierarchical models, *reference priors* in the form of flat priors and Jeffreys' priors, and some discussion of identifiability.

By a *reference prior*, we mean any prior that is not chosen for the information that it models. Rather, it is chosen to provide a common base for people to evaluate data. *We do not restrict our use of the term “reference prior” to the specific technical definition given by Bernardo (1979).* For those without access to scientific information, reference priors avoid the inconvenience of specifying informative prior distributions, so they are sometimes called *convenience priors*. People also talk about *noninformative priors*. That name stems from the fact that these priors typically have little influence on the posterior distribution. Although priors that have little effect on the posterior exist, there is no such thing as a noninformative prior. All priors express information about the parameters. Typically, the information expressed by “noninformative priors” is uniquely stupid in the sense that nobody would use them to make decisions in the absence of data. Their only possible virtue is that they have little affect on the posterior. Subsection 4.1.3 discusses a paradox in which the silly information from “noninformative” priors affects the posterior in a strange way.

Less mathematically inclined readers might want to proceed to Chapter 5 and return to this material as it arises in applications. Another possibility is to selectively read parts of this chapter now and return to other parts as needed.

### 4.1 Statistical Testing

The biggest difficulty in discussing testing is that people often overlook that there are two very different problems called testing. One problem is testing whether a specific model is correct. For example, we might claim that  $\theta = \theta_0$ , and want to know whether data are consistent or inconsistent with that claim. More precisely, we claim a distribution for the data  $f(y|\theta_0)$  and ask whether the data look like they could reasonably have come from that distribution. If not, we have a basis for questioning the claim. To evaluate whether observed data are consistent with an hypothesized model we use *p-values*. *P*-values are the probability of seeing data as weird or weirder than the actual data. They are computed using the one and only model available,  $f(y|\theta_0)$ . More recently, Bayesians have proposed *marginal* (Box, 1980) and *predictive p-values* (Rubin, 1984; Gelman et al. 2004) for the same purpose. Sometimes this problem is referred to as *significance testing*.

The other problem is a decision problem in which one must choose between two alternative models or hypotheses. This is the problem discussed by Bayesian methods as well as by the Neyman-Pearson theory of *hypothesis testing*. Bayesian testing is addressed by most books on Bayesian statistics. Significance testing and Neyman-Pearson theory are exposited, respectively, in classic books by Fisher (1956) and Lehmann (1986). Berger (2003) discussed whether the two approaches to hypothesis testing can be made to agree with each other as well as with *p*-values.

For those having a passing familiarity with the philosophy of science, a test of significance is a probabilistic method for falsifying an hypothesis. Neyman-Pearson testing and Bayesian testing are methods for evaluating the weight of evidence for the hypotheses. Bayesian analysis generally is a method for determining weights of evidence.

In this section we give a general discussion of testing, look at two approaches to checking Bayesian models, and examine the famous Lindley-Jeffreys paradox in Bayesian testing. We begin with an extensive example.

**EXAMPLE 4.1.1.** Consider two discrete densities indexed by  $\theta = 0, 1$  defined on the outcomes 2, 4, 6:

r	2	4	6
$f(r 0) \equiv \Pr(y = r \theta = 0)$	0.990	0.008	0.002
$f(r 1) \equiv \Pr(y = r \theta = 1)$	0.009	0.001	0.990
$LR \equiv f(r 0)/f(r 1)$	110	8	0.0091

Traditional significance testing can be used to address two separate issues: (1) do the data look like they could reasonably come from the  $f(r|0)$  model and (2) do the data look like they could reasonably come from the  $f(r|1)$  model. Both Neyman-Pearson and Bayesian testing address a different issue: given that the data come from one of these two models, which is correct? Significance testing asks whether a specific model seems reasonable whereas Neyman-Pearson and Bayes testing both provide a means of deciding between two models.

First, consider a significance test of  $f(r|0)$ . Under this model, there is a 99% chance of seeing  $y = 2$ . To see anything else would be extremely unusual and would cause us to doubt the model. In particular, we might reject the model if we see either  $y = 4$  or  $y = 6$ . The probability of rejecting the model, when it is true, is  $\alpha = 0.008 + 0.002 = 0.01$ . The  $p$ -value is defined to be the probability of seeing something as weird or weirder than we actually saw. “Weird” is defined by how unlikely it is to see particular data. Under  $f(r|0)$ , the weirdest thing we could see is  $y = 6$  because it has the smallest probability of occurring. The second weirdest thing is seeing  $y = 4$  because it has the second smallest probability of occurring. The  $p$ -value upon observing  $y = 6$  is  $p = 0.002$ , which is the probability of seeing  $y = 6$ . There is nothing weirder in this model than seeing  $y = 6$  (except seeing something other than a 2, 4, or 6). The  $p$ -value upon observing  $y = 4$  is the probability of seeing  $y = 4$  plus the probability of seeing data weirder than  $y = 4$ , which in this case means seeing  $y = 6$ . The  $p$ -value upon observing  $y = 4$  is  $p = \Pr(y = 4|\theta = 0) + \Pr(y = 6|\theta = 0) = 0.008 + 0.002 = 0.01$ .

A similar analysis occurs when performing a significance test of  $f(r|1)$ . Under this model, there is a 99% chance of seeing  $y = 6$ . Anything else is extremely unusual and causes us to doubt the model. We might reject the model if we see either  $y = 2$  or  $y = 4$  and the  $\alpha$  level would be  $\alpha = 0.009 + 0.001 = 0.01$ . The  $p$ -value is again the probability of seeing something as weird or weirder than we actually saw. Under  $f(r|1)$ , the weirdest thing we could see is  $y = 4$  because it has the smallest probability. The second weirdest observation is  $y = 2$ . The  $p$ -value upon observing  $y = 4$  is  $p = 0.001$ , which is the probability of seeing  $y = 4$ . There is nothing weirder in this model than seeing  $y = 4$ . The  $p$ -value upon observing  $y = 2$  is the probability of seeing  $y = 2$  plus the probability of seeing the weirder data,  $y = 4$ . The  $p$ -value upon observing  $y = 2$  is  $p = \Pr(y = 2|\theta = 1) + \Pr(y = 4|\theta = 1) = 0.009 + 0.001 = 0.01$ .

The significance test of  $f(r|0)$  has nothing to do with the test of  $f(r|1)$ . But seeing  $y = 4$  is unusual under both models and would make us doubt the validity of both models. In traditional significance testing, the model being tested is often called the null model. Sometimes the assumption of a valid null model is called the null hypothesis. There is no alternative hypothesis in traditional significance testing.

Neyman-Pearson testing and Bayes testing both involve a null hypothesis and an alternative hypothesis. In our example, we identify the null as  $H_0 : \theta = 0$  and the alternative as  $H_1 : \theta = 1$ .

Neyman-Pearson testing begins by picking a probability for rejecting the null hypothesis given that it is true. This is called the  $\alpha$  level. Since this obviously is the probability of making an error

(called a Type I error), we would like it to be as small as reasonable. Neyman-Pearson testing then seeks the best  $\alpha$  level test in the sense of maximizing the probability of rejecting the null hypothesis when the alternative is true. The probability of rejecting the null hypothesis when the alternative is true is called the *power* of the test, so Neyman-Pearson theory seeks the most powerful  $\alpha$  level test. Neyman-Pearson theory establishes that the most powerful  $\alpha$  level test is determined by rejecting  $H_0$  for the outcomes that have the smallest *likelihood ratios*  $f(r|0)/f(r|1)$ . Beginning with  $r$  values that have the smallest ratio, keep including  $r$  values until their probability computed under  $H_0$  is  $\alpha$ . In our example, for an  $\alpha = 0.01$  test, we first include the value  $r$  with the smallest  $f(r|0)/f(r|1)$ ,  $r = 6$ , with null probability 0.002 and then include  $r$  with the second smallest value of  $f(r|0)/f(r|1)$ ,  $r = 4$ , with null probability 0.008. The total probability under the null hypothesis is now 0.01, so the most powerful  $\alpha = 0.01$  level test rejects  $H_0$  when  $y$  is 4 or 6. The alert reader will notice that for an arbitrary  $\alpha$  level, perhaps  $\alpha = 0.02$ , some fudging (randomization) needs to take place.

In our example, both significance testing and Neyman-Pearson theory create an  $\alpha = 0.01$  test of the null hypothesis  $H_0 : \theta = 0$  that rejects the null hypothesis when  $y$  equals either 4 or 6. However, the reasons for using 4 and 6 are different. In significance testing, we reject when observing 4 and 6 because they have the smallest values of  $f(r|0)$  whereas in Neyman-Pearson testing, we reject because they have the smallest values of  $f(r|0)/f(r|1)$ . In other examples, significance tests and Neyman-Pearson tests may be different.

In significance testing, it is vital to have a very small  $\alpha$  level because the logic of the test is to reject the null model upon observing data that are unusual under the null model. If  $\alpha$  is not very small, we will reject for data that are not very unusual. We do not want to reject the null model when observing data that are actually consistent with it.

Although it is common practice in Neyman-Pearson testing to pick  $\alpha$  very small, there is no compelling reason to do so. In Neyman-Pearson testing, making  $\alpha$  very small decreases the power of the test. There is no compelling reason to have a very small  $\alpha$  at the expense of having small power. The appropriate procedure is to pick an  $\alpha$  that is reasonably small but that also gives a reasonably large power. This can be difficult to do. In fact, in our example, a very small  $\alpha$  is not small enough.

The Neyman-Pearson test in our example has a very small  $\alpha$  of 0.01 but gives unreasonable answers because  $\alpha$  is too large. In this example, the Neyman-Pearson  $\alpha = 0.01$  test rejects for  $y = 4$  or 6 and has a power of 0.991. However, the  $\alpha = 0.002$  test rejects for  $y = 6$  and has a power of 0.990. Increasing  $\alpha$  by 0.008 is not giving us a comparable increase in power, so the smaller test seems more desirable.

The  $\alpha = 0.01$  test rejects  $\theta = 0$  when  $y = 4$ , even though  $y = 4$  is 8 times more likely when  $\theta = 0$  than it is when  $\theta = 1$ , i.e.,  $f(4|0)/f(4|1) = 8$ . This seems like a clearly inappropriate decision barring outside information about either the prior probabilities of the two  $\theta$  values or the consequences of making mistakes. If you are forced to make a decision between  $\theta = 0$  and  $\theta = 1$  and see  $y = 4$ , other things being equal you would clearly pick  $\theta = 0$ . But the Neyman-Pearson  $\alpha = 0.01$  test rejects  $\theta = 0$ .

In the absence of information on the consequences of decisions, Bayesians would only reject  $\theta = 0$  when observing  $y = 4$  if they previously had an initial probability for  $\theta = 1$  that was at least 8 times greater than the probability of  $\theta = 0$ . The Bayesian bases decisions on the posterior probabilities for  $\theta$ . In particular, for prior probabilities  $p(0)$  and  $p(1)$ , a Bayesian computes

$$p(0|4) \equiv \Pr(\theta = 0|y = 4) = \frac{f(4|0)p(0)}{f(4|0)p(0) + f(4|1)p(1)} = \frac{0.008p(0)}{0.008p(0) + 0.001p(1)}$$

and bases a decision on this posterior probability. If you have equal prior probabilities on the two models, i.e.,  $p(0) = 0.5$ , then after the data you have  $p(0|4) = 8/9$ , which is pretty strong evidence for model  $H_0$ . On the other hand, it takes a prior 8 times higher for the model under  $H_1$  in order for the data  $y = 4$  to make you indifferent between the models, that is, for  $p(0) = 1/9$ , so that  $p(1) = 8p(0)$ , Bayes' Theorem gives  $p(0|4) = 0.5$ . We think the Bayesian has the much better process for deciding between  $\theta = 0$  and  $\theta = 1$ .

Incidentally, our two significance tests call both models in question when observing  $y = 4$ , which is probably the most intelligent response to seeing  $y = 4$ . Bayesian methods give good results within the context of the assumed models but do not naturally provide for questioning the models. Significance testing is all about questioning the model. See Christensen (2005, 2008) for more detailed discussions on the alternative methods of testing. See Subsections 1 and 2 for checking Bayesian models.

Just as high prior probability on  $\theta = 1$  can affect our response to observing  $y = 4$ , higher prior probability on  $\theta = 1$  can make us unsure about  $\theta$  even when observing  $y = 2$ . Normally, observing  $y = 2$  clearly indicates that  $\theta$  is 0 rather than 1, but if  $p(1) = 110p(0)$ , upon seeing  $y = 2$  we would be indifferent between the hypotheses.

### *Parametric Testing*

In practice, hypothesis testing frequently involves some function of the parameter vector, say,  $\gamma(\theta)$ , and either one-sided null and alternative hypotheses like  $H_0 : \gamma(\theta) \leq \gamma_0$  versus  $H_1 : \gamma(\theta) > \gamma_0$ , or point null and two-sided alternatives like  $H_0 : \gamma(\theta) = \gamma_0$  versus  $H_1 : \gamma(\theta) \neq \gamma_0$ . The function  $\gamma(\theta)$  might be the first component of  $\theta$ , say,  $\gamma(\theta) = \theta_1$ , or it might be the difference of the first two components,  $\gamma(\theta) = \theta_1 - \theta_2$ , or the ratio  $\gamma(\theta) = \theta_1/\theta_2$ . Pick  $\gamma(\theta)$  to be appropriate for the specific problem at hand.

To test  $H_0 : \gamma(\theta) \leq \gamma_0$  versus  $H_1 : \gamma(\theta) > \gamma_0$ , we need to find the marginal posterior distribution of  $\gamma \equiv \gamma(\theta)$ , either analytically or by simulation. Given the marginal posterior, simply find  $\Pr[\gamma \leq \gamma_0 | y]$  and decide in favor of  $H_0$  if  $\Pr[\gamma \leq \gamma_0 | y]$  is sufficiently large. If forced to make a decision, in the absence of information on the consequences of decisions, one would pick whichever hypothesis had posterior probability greater than 0.5. Recall that  $\Pr[\gamma \leq \gamma_0 | y] = 1 - \Pr[\gamma > \gamma_0 | y]$ . In practice, unless the posterior probabilities are close to 0 and 1, it may be better to admit that we do not know which hypothesis is true.

Testing  $H_0 : \gamma = \gamma_0$  versus  $H_1 : \gamma \neq \gamma_0$  is more difficult. With a continuous prior on  $\theta$ , typically  $\Pr(\gamma = \gamma_0) = 0$ . To make an interesting testing problem, there must be positive prior probability that the null hypothesis is true. Define  $q_0 \equiv \Pr[\gamma = \gamma_0] > 0$  as this prior probability. We then require conditional prior distributions for  $\theta$  under each of the hypotheses, that is,  $p(\theta | \gamma = \gamma_0) \equiv p_0(\theta)$  and  $p(\theta | \gamma \neq \gamma_0) \equiv p_1(\theta)$ . These “densities” have some mathematical problems. The easier one is  $p_1(\theta)$ . We could specify an overall continuous prior for  $\theta$ , say  $p_*(\theta)$  and, since  $p_1$  is obtained by conditioning on an event that occurs with prior probability one under  $p_*$ , just take  $p_1 = p_*$ . The more difficult conditional distribution is  $p_0(\theta)$  because it is conditioned on  $\gamma = \gamma_0$ . For example, if  $\gamma$  is defined by  $\gamma(\theta) = \theta_1$ , then we need the conditional distribution of  $\theta_2, \dots, \theta_r$  given that  $\theta_1 = \gamma_0$ . If  $\gamma$  is defined by  $\gamma(\theta) = \theta_1 - \theta_2$ , then we need the conditional distribution of, say,  $\theta_1 + \theta_2, \theta_3, \dots, \theta_r$  given that  $\theta_1 - \theta_2 = \gamma_0$ . In theory, these could also be obtained from an overall prior  $p_*$ .

With these pieces in place, the prior on  $\theta$  is defined by

$$p(\theta) = q_0 p_0(\theta) I_{\{\gamma_0\}}(\gamma(\theta)) + (1 - q_0) p_1(\theta) [1 - I_{\{\gamma_0\}}(\gamma(\theta))].$$

It is then easy to show that

$$\Pr(\gamma = \gamma_0 | y) = \frac{q_0 f_0(y)}{q_0 f_0(y) + (1 - q_0) f_1(y)}$$

where  $f_1(y) = \int f(y | \theta) p_1(\theta) d\theta$  and  $f_0(y) = \int f(y | \theta) p_0(\theta) d\theta$ . In  $f_0$  we are being a little loose with the notation used in the integral because the integration and the density  $p_0(\theta)$  are only defined for  $\theta$  values with  $\gamma(\theta) = \gamma_0$ . (It would be a simple matter to define the integral rigorously using measure theory.) Fundamentally, this testing problem is no more difficult than Example 4.1.1. It is a simple test of two sampling densities:  $f_0(y)$  and  $f_1(y)$ . The difference, and the difficulty, is that these sampling densities are obtained by averaging  $f(y | \theta)$  over the conditional priors  $p_0(\theta)$  and

$p_1(\theta)$ . This argument also applies to the one-sided hypotheses, but the analysis discussed earlier is simpler. A slightly more general version of Bayesian parameter testing is given in Subsection 4.8.1.

The next two exercises should make you a tougher, if not necessarily better, statistician.

**EXERCISE 4.1.** Let  $y|\theta \sim \text{Pois}(\theta)$  and assume a prior on  $\theta$  of  $\text{Gamma}(1,1)$ . (a) For  $H_0 : \theta \leq 1$  versus  $H_1 : \theta > 1$  obtain the formula for the posterior probability that  $H_0$  is true. Calculate the probability using WinBUGS for  $y = 3, 5$ , and  $7$ . (b) For  $H_0 : \theta = 1$  versus  $H_1 : \theta \neq 1$  with  $q_0 = 0.5$  and  $p_1(\theta) = e^{-\theta}$ , obtain the analytical formula for the posterior probability that  $H_0$  is true. Use R to calculate the exact probabilities for the three  $y$  values. You should get  $0.4952$ ,  $0.1640$ , and  $0.0183$ , respectively. (c) Use the following WinBUGS code to obtain numerical approximations to the exact results obtained in R by monitoring  $z$ . Explain why the WinBUGS code is doing the appropriate thing.

```
model{
  y ~ dpois(theta)
  z ~ dbern(q0)
  theta0 <- 1
  theta1 ~ dexp(1)
  theta <- z*theta0 + (1-z)*theta1
}
list(y=3, q0=0.5)
list(theta1=1,z=1)
```

**EXERCISE 4.2.** Let  $y_1|\theta_1 \sim \text{Exp}(\theta_1)$  and  $y_2|\theta_2 \sim \text{Exp}(\theta_2)$  with  $y_1 \perp\!\!\!\perp y_2 | \theta_1, \theta_2$ . For  $H_0 : \theta_1 - \theta_2 = 0$  versus  $H_1 : \theta_1 - \theta_2 \neq 0$  use  $q_0 = 0.5$ , the conditional prior  $p_1(\theta_1, \theta_2) = e^{-(\theta_1+\theta_2)}$ , and with  $\theta_1 = \theta_2 = \theta$ ,  $p_0(\theta) = e^{-\theta}$ . (a) Derive the joint posterior under  $H_1$ . (b) Derive the posterior under  $H_0$ . (c) Derive the explicit formula for the posterior probability of  $H_0$ . (d) If  $y_1 = 500$  and  $y_2 = 2$ , obtain the posterior probability of  $H_0$  both analytically using R and also using the WinBUGS code below.

```
model{
  for(i in 1:2){
    y[i] ~ dexp(theta[i])
    ttheta[i] ~ dexp(1) # priors on distinct thetas
    theta[i] <- z*mu + (1-z)*ttheta[i] # z indicates common
    } # versus different rates
  z ~ dbern(q0)
  mu ~ dexp(1) # mu is the common rate
}
list(y=c(500,2), q0=0.5)
list(z=1,ttheta=c(1,1),mu=1)
```

In Exercises 4.1 and 4.2, monitoring the simulated  $\theta$ s would not be useful since some of the values are taken under  $H_0$  and others are taken under  $H_1$ . If, in a given example, the posterior probability of  $H_i$  were very large, then you could re-run an analysis under that model to approximate the posterior.

#### 4.1.1 Checking Bayesian Models

Box (1980) suggested that Bayesians should use significance testing based on the marginal density of the data  $f(y)$  as a method of model checking. In this method, for observed data  $y_{obs}$ , the model is rejected if  $y_{obs}$  looks too implausible under the presumed model. Specifically, the model is rejected

if the density of our observed data  $f(y_{obs})$  is too small. “Too small” is quantified by finding the  $p$ -value, the probability of seeing data with a density as small or smaller than  $y_{obs}$ ,

$$p = \Pr[f(y) \leq f(y_{obs})] = \int I_{(-\infty, f(y_{obs})]}[f(y)]f(y)dy. \quad (1)$$

The  $p$ -value gives the probability of seeing data that are no more plausible than we actually saw. If  $p$  is “too small” we regard  $y_{obs}$  as insufficiently plausible to support the presumed model.

**EXAMPLE 4.1.2. Model Checking a Binomial.** At the end of Example 2.3.1 we showed that for binomial data with a beta prior, the predictive distribution is a beta-binomial. Similarly, the marginal distribution is also a beta-binomial. Suppose  $y$  is distributed as a binomial random variable with  $n$  trials and probability  $\theta$ , that is,

$$y|\theta \sim \text{Bin}(n, \theta).$$

The sampling density is

$$f(y|\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}.$$

Take the prior to be  $\theta \sim \text{Beta}(a, b)$ , so

$$p(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1}.$$

Using the sampling density for  $y$  and the prior on  $\theta$  in strict analogy to how, in Example 2.3.1, we used the sampling distribution of future observations  $\tilde{y}$  and the posterior on  $\theta$ , we arrive at the marginal distribution of  $y$  as

$$f(y) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \binom{n}{y} \frac{\Gamma(y+a)\Gamma(n-y+b)}{\Gamma(n+a+b)}$$

for  $y = 0, 1, \dots, n$ . This is again a Beta-Binomial distribution.

In particular, suppose we believe there is a very small prior probability of success, say  $\theta \sim \text{Beta}(1, 7)$ . This has a prior mean of  $1/(1+7) = 0.125$ . Furthermore, suppose we have  $n = 8$  trials. Using the formula for the Beta-Binomial, some values of the marginal distribution are  $f(0) = 7/15 = 0.47$ ,  $f(8) = 1/6435 = 0.000155$ , and  $f(7) = (7/4)f(8) = 0.000272$ . A significance test identifies  $y$  values that have very small values of  $f(y)$  and rejects the hypothesized model if one of these rare  $y$ s actually occurs. Here,  $f(0)$  is not small, but  $f(7)$  and  $f(8)$  are small. To do an  $\alpha = 0.000427$  level test, we reject if  $y = 7$  or  $8$ . So if we actually got 7 successes in this problem, we could be reasonably certain that something was wrong with our model. How could we be so unlucky as to observe data that are so unlikely if our model were actually true? Rejecting the model does not tell us what is wrong. Perhaps the data are not binomial. Perhaps the trials are not independent. Perhaps the prior is inappropriate. The test only suggests (quite strongly) that *something* about the assumptions we made is incorrect. It is now our task to identify which assumptions are unreasonable.

**EXERCISE 4.3.** How do the results of the significance test change if the prior is  $\text{Beta}(0.1, 0.7)$ ? What if we use  $\text{Beta}(10, 70)$ ? Note that all three prior distributions have the same mean.

More generally, one could look at any discrepancy statistic  $D = D(y)$ , for example,

$$D(y) = \sum_{i=1}^n \frac{[y_i - E(y_i)]^2}{\text{Var}(y_i)},$$

with the mean and variance computed from the marginal distribution. Then we find  $D$ 's marginal density  $h(d)$  and compute the  $p$ -value

$$p = \Pr[h(D) \leq h(D(y_{obs}))].$$

One can also evaluate individual observations  $y_i$ , cf. Subsection 8.3.1.

**EXERCISE 4.4.** Suppose in a random sample of 10 transportation workers, all were found to be on drugs. Calculate the marginal  $p$ -value (1) to evaluate whether such data are consistent with a model that has

$$y_1, \dots, y_{10} | \theta \stackrel{iid}{\sim} \text{Bern}(0.10).$$

(Think of this as having a prior with  $\Pr[\theta = 0.1] = 1.0$ .) Would the test change if it were based on  $\sum_{i=1}^{10} y_i \sim \text{Bin}(10, 0.10)$ ?

**EXERCISE 4.5.** Let  $y_i | \theta \stackrel{iid}{\sim} \text{Exp}(\theta)$ ,  $i = 1, 2$ , and let  $p(\theta) = e^{-\theta}$ . Suppose  $(y_1, y_2) = (5, 8)$  is observed. Calculate the marginal  $p$ -value from (1) and decide whether the data are inconsistent with the model. You may use WinBUGS or R to assist in answering the question.

#### 4.1.2 Predictive P-Values

Gelman, Meng, and Stern (1996) proposed a method of assessing the fit of a model that is based on predictive observations. Consider a future data set  $\tilde{y}$  that comes from the same model as the current data  $y$ , that is,  $\tilde{y} \perp\!\!\!\perp y | \theta$  with  $f(y|\theta)$  and  $f(\tilde{y}|\theta)$  having the same functional form. Define a discrepancy  $D(y; \theta)$  that measures the fit of the observed data to the presumed sampling model. For example, with a sample of conditionally iid data of size  $n$ , partition the real line into  $m$  sets (bins) with  $n_i$  of the  $y_j$ s falling into set  $i$ . The Pearson chi-square discrepancy is

$$D_P(y; \theta) = \sum_{i=1}^m \frac{[n_i - E(n_i|\theta)]^2}{E(n_i|\theta)},$$

where  $E(n_i|\theta)$  is  $n$  times the probability of  $y_j$  falling in the  $i$ th bin given  $\theta$ .

Gelman, Meng, and Stern (1996) recommend calculating the “predictive p-value”

$$\begin{aligned} ppv &\equiv \Pr[D(\tilde{y}; \theta) \geq D(y; \theta) | y] \\ &= \int \int I_{[D(\tilde{y}; \theta) \geq D(y; \theta)]}(\tilde{y}, \theta) f(\tilde{y} | \theta) p(\theta | y) d\tilde{y} d\theta. \end{aligned} \quad (2)$$

More formally, with  $A = \{(u, v) : D(u; v) \geq D(y; v)\}$ ,

$$ppv = \Pr(A | y) = \int \int I_A(\tilde{y}, \theta) f(\tilde{y} | \theta) p(\theta | y) d\tilde{y} d\theta.$$

This is numerically approximated by sampling  $\theta^k$  from the posterior of  $\theta$  followed by sampling  $y^k$  from  $f(\tilde{y} | \theta^k)$  and approximating

$$ppv \doteq \sum_{k=1}^s I[D(y^k; \theta^k) > D(y; \theta^k)]/s.$$

Unlike Box's procedure, this is not a traditional significance test. Nonetheless, the recommendation is that if  $ppv$  is very small or very large, the observed data are inconsistent with the assumed model. There are many possible choices of discrepancy measure such as, for a random sample,

$$D(y; \theta) = \sum_{i=1}^n \frac{[y_i - E(y_i|\theta)]^2}{\text{Var}(y_i|\theta)}.$$

If the discrepancy measure does not depend on  $\theta$ , for example

$$D(y) = \sum_{i=1}^n \frac{[y_i - E(y_i)]^2}{\text{Var}(y_i)},$$

with the mean and variance computed from the marginal distribution, then evaluation of the *ppv* is made relative to the predictive distribution (but it is still not a significance test). See Gelman, Meng, and Rubin (1996) for more details.

#### 4.1.3 Lindley-Jeffreys Paradox

The moral of the Lindley-Jeffreys paradox is that if you pick a stupid prior, you can get a stupid posterior. Suppose  $y|\theta \sim N(\theta, 1)$  and we want to test the hypothesis  $H_0 : \theta = 0$  versus the alternative  $H_1 : \theta > 0$ . We need to specify a prior probability distribution on the set of all possible  $\theta$  values, i.e., on  $\theta \geq 0$ . As previously discussed, if we use a continuous distribution on  $\theta \geq 0$ , the prior probability of  $\theta = 0$  is 0, so there will be no chance of ever accepting the null hypothesis. We thus specify  $q_0 = \Pr(\theta = 0) = 0.5$ , and distribute the rest of the probability on  $\theta > 0$ . To distribute the rest of the probability we specify that  $1 - q_0 = \Pr(\theta > 0) = 0.5$  and further specify a continuous distribution for  $\theta$  given  $\theta > 0$ .

To make our point, we need a distribution that is highly dispersed. Such distributions are often thought to convey ignorance about the particular value of  $\theta > 0$ , but we have reservations about using such models to convey ignorance. For no other reason than to simplify computations, we use a normal distribution as an approximate conditional distribution. In particular, we approximate  $\theta | \theta > 0$  with a  $N(\mu_0, \sigma_0^2)$ . To have high dispersion, we need  $\sigma_0^2$  large, and to make this a reasonable approximation,  $\mu_0$  needs to be considerably larger than  $\sigma_0$  so that there is very little probability of observing  $\theta < 0$ .

Example 2.3.3 gave a Bayesian analysis for normal data with known variance and a normal prior. Computing the marginal distribution of  $y$  in that case would give the approximation

$$y | \theta > 0 \sim N(\mu_0, 1 + \sigma_0^2).$$

For example, we might take  $\sigma_0^2 = e^7 \doteq 1,097$  and  $\mu_0 = e^9 \doteq 8,103$ .

Given data  $y$ , to decide between the hypotheses we calculate the posterior probability of  $H_0$ :

$$\Pr(\theta = 0 | y) = \frac{q_0 f(y | \theta = 0)}{q_0 f(y | \theta = 0) + (1 - q_0) f(y | \theta > 0)}.$$

With  $q_0 = 0.5$ , it is not difficult to see that  $\Pr(\theta = 0 | y) > 0.5$  if and only if  $f(y | \theta = 0) > f(y | \theta > 0)$ . These are both normal densities:

$$f(y | \theta = 0) = \frac{1}{\sqrt{2\pi}} \exp[-y^2/2]$$

and

$$f(y | \theta > 0) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{1 + \sigma_0^2}} \exp[-(y - \mu_0)^2 / 2(\sigma_0^2 + 1)].$$

Note that  $f(y | \theta = 0) > f(y | \theta > 0)$  if and only if

$$y^2 < \log(1 + \sigma_0^2) + (y - \mu_0)^2 / (\sigma_0^2 + 1).$$

If we take  $\sigma_0^2 = e^7$ ,  $\mu_0 = e^9$ , and  $y \leq \sqrt{7} \doteq 2.646$ , then

$$y^2 \leq 7 < \log(1 + \sigma_0^2) < \log(1 + \sigma_0^2) + (y - \mu_0)^2 / (\sigma_0^2 + 1)$$

so we will choose  $\theta = 0$  rather than  $\theta > 0$ .

The paradoxical thing about this is that seeing  $y = 2.6$  from a  $N(0, 1)$  distribution is an exceptionally unusual event. In significance testing, one would easily reject the model with  $\theta = 0$  when seeing  $y = 2.6$ . The problem here is that the prior distribution is putting so much probability on very large values of  $\theta$  that, even though it is very unlikely that one would see  $y = 2.6$  from a  $N(0, 1)$ , it is more likely to come from the  $N(0, 1)$  distribution than it is from the distribution of  $y$  given the alternative, namely, the  $N(e^9, 1 + e^7)$  distribution. Note that similar computations hold when the distribution of  $\theta$  given the alternative is a  $N(e^{19}, e^{17})$  distribution, so that we would accept  $\theta = 0$  even if we saw  $y = \sqrt{17} \doteq 4.1$ .

We believe that there is no such thing as a prior distribution that embodies ignorance; however, there are some distributions that are more easily overwhelmed by the data than others, and these are often prior distributions with large variability. In any case, the moral of the Lindley-Jeffreys Paradox is that you need to actually think about an appropriate distribution for  $\theta$  given the alternative. For this problem, trying to choose a convenient distribution containing little information wreaks havoc.

## 4.2 Exchangeability

The use of parameters is a convenience, not a fundamental aspect of Bayesian analysis. Examples 2.3.2, 2.3.3, and much statistical work of all kinds, focus on observations that are iid given a parameter. Bayesians then incorporate a prior distribution on the parameter.

Rather than focusing on iid observations given a parameter, a more fundamental concept is that of *exchangeable* random variables. To be exchangeable, the  $y_i$ s must have a (marginal) distribution for the vector  $y$  that is the same regardless of the order in which the observations are written down. The Bayesian conditional iid formulation satisfies this as will be illustrated in Example 4.2.1. In the broader context, specifying parameters serves merely as a convenient method of attaining exchangeability.

A famous representation theorem by de Finetti establishes that exchangeable random variables always can be generated by introducing some parameter and using the Bayesian conditional iid formulation. To deemphasize the role of parameters in Bayesian analysis, an alternative mindset must be developed such as focusing on prediction of future observable values rather than estimating unknowable parameters. This is discussed further in Section 4.5.

Although exchangeable random variables can be written as conditionally independent given a parameter, they are not typically unconditionally independent. This distinction is examined in the next example.

EXAMPLE 4.2.1. *Bernoulli Data.* In Example 2.3.2 we assumed

$$y_1, \dots, y_n | \theta \stackrel{iid}{\sim} \text{Bern}(\theta)$$

with

$$\theta \sim \text{Beta}(a, b).$$

The  $y_i$ s are modeled as being conditionally independent given  $\theta$ , but they are not unconditionally independent. We can see this by using a well known result on conditional distributions (Proposition B.2) to show that their covariances are not 0.

$$\begin{aligned} \text{Cov}(y_i, y_j) &= \text{Cov}[\text{E}(y_i | \theta), \text{E}(y_j | \theta)] + \text{E}[\text{Cov}(y_i, y_j | \theta)] \\ &= \text{Cov}[\theta, \theta] + \text{E}[0] \\ &= \text{Var}(\theta) \\ &= \frac{\mu_0(1 - \mu_0)}{a + b + 1} > 0, \quad \mu_0 \equiv a/(a + b). \end{aligned}$$

Here the last equality is a well known fact about Beta distributions, cf. Table 2.2. Note that the covariance does not depend on the choices of  $i$  or  $j$ .

EXERCISE 4.6. Show that, unconditionally, the  $y_i$ s in Example 4.2.1 are still identically distributed Bernoulli random variables by finding the unconditional probability that  $y_i$  equals 1.

The joint density of  $\theta$  and  $y$  is  $p(\theta, y) = f(y|\theta)p(\theta)$  and, integrating out  $\theta$ , the marginal density of  $y$  is

$$f(y) = \int f(y|\theta)p(\theta)d\theta.$$

In the conditional iid case of Section 2.3, we get exchangeability. The conditional density for  $y$  is

$$f(y|\theta) = \prod_{i=1}^n f_*(y_i|\theta),$$

in which order clearly does not matter, so in computing  $f(y)$  order also does not matter.

EXAMPLE 4.2.2. *Conditional and Unconditional Independence.* Suppose in a random sample of 1,000 transportation workers, all were found to be on drugs. You are asked to predict the outcome of the 1,001st individual who will now be sampled. Obviously, you would predict that the new person would be on drugs. Clearly, we think that the past observations contain information about the future observation, so the future is not independent of the past. Typically, all of these observations would be modeled as independent conditional on the probability of drug use  $\theta$ . The past gives information about  $\theta$  and that information is used to inform our prediction about the future observation. In a Bayesian analysis,  $\theta$  is treated as a random variable, so if we integrate it out to look at the marginal distribution of the observations, they are unconditionally dependent. Unconditional dependence is proper because we think that the past should inform about the future.

In non-Bayesian statistical analyses,  $\theta$  is treated as fixed. If you were told that exactly 10% of workers in the transportation industry are on drugs and you truly believe the model, the future observation will be independent of the past and there is only a 10% chance of a randomly sampled person being on drugs, so you would predict that the next person will not be on drugs, in spite of the data in which everyone is on drugs. With  $\theta$  known to be 10% and random sampling, the future is independent of the past, so the data are irrelevant.

Of course the data here are clearly inconsistent with the model of 10% on drugs, so something is clearly wrong (see Exercise 4.4). Either the observations are not (conditionally) independent observations from the population or the proportion of drug use is not 10%. As discussed in Section 1, the basis of statistical significance tests is checking whether data are consistent with the model being considered.

Finally, with  $\theta$  treated as fixed but unknown, the past would be used to inform us about the unknown constant  $\theta$  and this information would be used in predicting a new observation. An obvious estimate for these data is  $\hat{\theta} = 1$  with corresponding prediction that a new person is *certainly* on drugs. This involves treating  $\hat{\theta}$  like it is really  $\theta$ , which it is not. For known  $\theta$ , it is appropriate to use  $f(y|\theta)$  to make predictions but using  $f(y|\hat{\theta})$  to make predictions ignores the fact that  $\theta$  is unknown and being estimated. Predictions based on  $f(y|\theta)$  would be conditionally independent of the past. The prediction using  $\theta = \hat{\theta}$  only seems to be conditionally independent of the past. Moreover, non-Bayesians make no specification that would allow an unconditional evaluation of the dependence between the past and the future.

Any reasonable Bayesian analysis would retain some uncertainty in the prediction. Bayesians would never predict that the next person would *certainly* be on drugs. For example, with a Beta(1,1) [uniform] prior, the predictive probability of being on drugs is  $(1,001)/(1,002) < 1$ . The Bayesian approach deals directly with all of these issues.

EXERCISE 4.7. Let  $y_i|\theta \stackrel{iid}{\sim} \text{Bernoulli}(\theta)$ , and let  $p(\theta) = I_{[0,1]}(\theta)$ , i.e.,  $\theta$  is  $U[0,1]$ . (a) Obtain the marginal density of  $(y_1, \dots, y_n)$ . (b) Calculate the predictive probability that  $y_{n+1} = 1$  given that  $y_1 = \dots = y_n = 1$ . Simplify the formula you get using the fact that  $\Gamma(a+1) = a\Gamma(a)$  and thus establish that the answer is 1,001/1,002 if  $n = 1,000$ .

### 4.3 Likelihood Functions

When viewed as a sampling density,  $f(y|\theta)$  is a function of  $y$  for fixed  $\theta$ . The density provides a specification of the probabilities for various possible data. In statistics, we eventually get to see the data, say  $y = y_{obs}$ , and we want to draw inferences (conclusions) about  $\theta$ . Thus, we are interested in the values of  $\theta$  that are most likely to have generated  $y_{obs}$ . Such information comes from  $f(y_{obs}|\theta)$  but with  $y_{obs}$  fixed and  $\theta$  allowed to vary. This new way of thinking about  $y$  and  $\theta$  determines a new function called the *likelihood function*, written

$$L(\theta|y_{obs}) \equiv f(y_{obs}|\theta).$$

The likelihood function and the sampling density are different concepts based on the same collection of symbols. As a sampling density  $f(y|\theta)$ ,  $\theta$  is fixed and  $y$  is the variable. As a likelihood  $L(\theta|y)$ ,  $y$  is fixed and  $\theta$  is allowed to vary. In the conditionally iid case we assume that, given the parameter vector  $\theta$ , the random variables  $y_i$  are iid with density  $f_*(\cdot|\theta)$ , so the likelihood function is

$$L(\theta|y) = \prod_{i=1}^n f_*(y_i|\theta).$$

A standard non-Bayesian estimate of  $\theta$  uses the value of  $\theta$  that maximizes the likelihood. The value  $\hat{\theta} \equiv \hat{\theta}(y)$  with

$$L(\hat{\theta}|y) = \sup_{\theta} L(\theta|y)$$

is called the *maximum likelihood estimate (MLE)* of  $\theta$ .

**EXAMPLE 4.3.1.** Consider again the two discrete densities defined in Example 4.1.1:

$r$	2	4	6
$f(r 0)$	0.990	0.008	0.002
$f(r 1)$	0.009	0.001	0.990

We view these as the only two distributions that are possible for our data, but we do not know which is appropriate. In other words, we want to use the data to decide which of the parameter values is more likely. The likelihood function only takes on two values,  $L(0|y)$  and  $L(1|y)$ . In this case, the answers are very clear. If we observe  $y = 2$ ,  $\theta = 0$  is far more likely than  $\theta = 1$ ; moreover,  $\hat{\theta}(2) = 0$ . If we see  $y = 6$ ,  $\theta = 1$  is far more likely, and if we see  $y = 4$ ,  $\theta = 0$  is somewhat more likely. This is exactly the intuition we used to call in question the appropriateness of Neyman-Pearson testing.

Typically, the likelihood function is only of interest up to constants of proportionality. It is the shape of the likelihood function, not the actual values that are important. Multiplying by a constant (any number that does not depend on  $\theta$ ) is irrelevant. Proportional likelihoods give the same MLE. (But frequentist confidence intervals and tests can differ for proportional likelihoods.)

**EXAMPLE 4.3.2.** Suppose that 10 workers were sampled and that two of them tested positive for drug use. The likelihood is  $L(\theta|y = 2) \propto \theta^2(1 - \theta)^8$ . A plot is given in Figure 4.1. Both  $\theta = 0$  and 1 are impossible, since they exclude the possibility of seeing drug tests that are both positive and negative. Values of  $\theta$  above 0.5 are particularly unlikely to have generated these data since far less than half the observations were positive. The MLE for  $\theta$  is the sample proportion of positives,  $0.20 = 2/10$ . Although it takes a bit of calculus to prove it, the likelihood function in Figure 4.1 clearly has a maximum near 0.2.

In Bayes' Theorem,

$$p(\theta|y) = \frac{L(\theta|y)p(\theta)}{\int L(\theta|y)p(\theta)d\theta}.$$

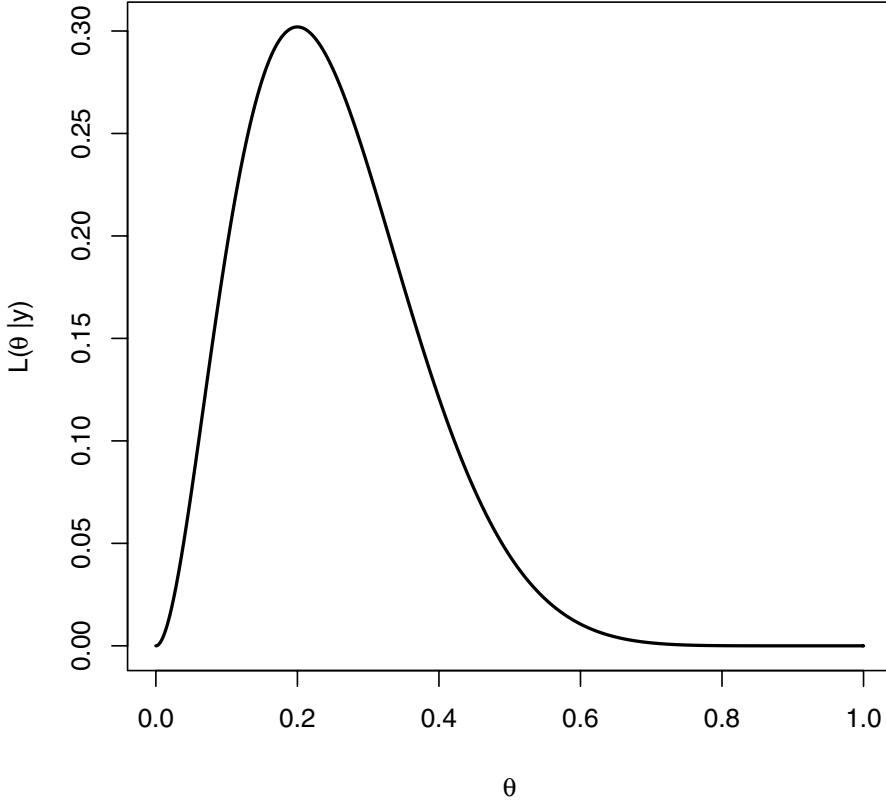


Figure 4.1: Likelihood for  $\text{Bin}(10, \theta)$  with  $y = 2$ .

Considering this form, any constants of proportionality, that is, multiplicative terms in the likelihood function that do not involve  $\theta$ , cancel out between the numerator and denominator, so are irrelevant.

To make this more specific, suppose we conduct two experiments to learn about  $\theta$ . These give two likelihoods  $L_1(\theta|y_1)$  and  $L_2(\theta|y_2)$  based on different data  $y_1$  and  $y_2$ . Suppose further that the likelihoods are proportional, that is,

$$L_1(\theta|y_1) = K L_2(\theta|y_2),$$

where  $K$  is a constant of proportionality. It is a simple matter to see that for either data  $y_i$ ,

$$p(\theta|y_i) = \frac{L_1(\theta|y_1)p(\theta)}{\int L_1(\theta|y_1)p(\theta)d\theta} = \frac{L_2(\theta|y_2)p(\theta)}{\int L_2(\theta|y_2)p(\theta)d\theta}.$$

The posteriors are the same, so all Bayesian inferences are the same.

An interesting example of proportional likelihoods comes from Bernoulli trials. We will see that a Bayesian only cares about the number of trials and the number of successes. A Bayesian does not care whether the data came about from fixing the number of trials and observing the random number of successes (binomial sampling) or from fixing the number of successes and observing the random number of trials needed to obtain them (negative binomial sampling).

**EXAMPLE 4.3.3.** *Bernoulli Trials.* Both the binomial distribution and the negative binomial distribution involve a sequence of independent success-failure trials in which each trial has probability of success  $\theta$ . The difference between the binomial and the negative binomial is in how the sequence terminates, i.e., their *stopping rules*. In a binomial, one predetermines a number of trials  $n$  and the random variable  $y_1$  is the number of successes. Thus the number of successes is  $s = y_1$  and the number of failures is  $f = n - y_1$ . In a negative binomial, one predetermines a number of successes  $s$  and the random variable  $y_2$  is the number of trials required to get  $s$  successes. In the negative binomial  $s$  is specified and  $f = y_2 - s$ .

In Example 2.3.1 we established that when

$$y_1|\theta \sim \text{Bin}(n, \theta)$$

with sampling density and likelihood

$$f_1(y_1|\theta) = L_1(\theta|y_1) = \binom{n}{y_1} \theta^{y_1} (1-\theta)^{n-y_1}$$

and conjugate prior

$$\theta \sim \text{Beta}(a, b),$$

the posterior distribution is

$$\theta|y_1 \sim \text{Beta}(y_1 + a, n - y_1 + b) = \text{Beta}(s + a, f + b). \quad (1)$$

Now suppose  $y_2|\theta$  is distributed negative binomial with  $s$  successes. The sampling density and likelihood are

$$f_2(y_2|\theta) = L_2(\theta|y_2) = \binom{y_2 - 1}{s - 1} \theta^s (1-\theta)^{y_2 - s}.$$

If  $s = y_1$  and  $n = y_2$ , then we have proportional likelihoods

$$L_1(\theta|y_1) = \left[ \binom{n}{y_1} / \binom{y_2 - 1}{s - 1} \right] L_2(\theta|y_2).$$

For this to occur, the random number of successes in the binomial must equal the fixed number of successes in the negative binomial and the fixed number of trials in the binomial must equal the random number of trials in the negative binomial.

Again using the  $\theta \sim \text{Beta}(a, b)$  prior distribution and applying Bayes' Theorem to the negative binomial data

$$\begin{aligned} p(\theta|y_2) &= \frac{\binom{y_2 - 1}{s - 1} \theta^s (1-\theta)^{y_2 - s} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1}}{\int \binom{y_2 - 1}{s - 1} \theta^s (1-\theta)^{y_2 - s} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1} d\theta} \\ &= \frac{\theta^{s+a-1} (1-\theta)^{y_2 - s + b - 1}}{\int \theta^{s+a-1} (1-\theta)^{y_2 - s + b - 1} d\theta}, \end{aligned}$$

which, as in Example 2.3.1 we recognize as a beta density, so

$$\theta|y_2 \sim \text{Beta}(s + a, y_2 - s + b) = \text{Beta}(s + a, f + b). \quad (2)$$

The posteriors (1) and (2) agree. The posterior depends on the prior as well as the numbers of successes and failures, but it does not depend on the stopping rule. In other words, in a sequence of Bernoulli trials, it does not depend on whether we decide to stop after  $n$  trials or if we stop when we obtain  $s$  successes.

The *stopping rule principle* is quite general. It is that the analysis of a sequence of observations should not depend on the stopping rule. The Bayesian analysis depends on the actual results observed from a sequence of observations but it does not depend on how the sequence of observations was stopped, so the Bayesian analysis satisfies the stopping rule principle.

The idea that constants of proportionality in the likelihood function are irrelevant is a fundamental idea. The idea that whenever two likelihoods are proportional, all statistical inference should be identical is known as the *Likelihood Principal*. Bayesian analysis satisfies the likelihood principle. Berger and Wolpert (1984) have written a wonderful book on the likelihood principle. Likelihood functions are revisited in Section 4.13.

In analyzing data, particularly measurement data, one frequently transforms the data prior to the analysis. Thus we analyze, say, the logs or the square roots rather than the original data. The next exercise establishes that *the likelihood, and therefore the Bayesian analysis, does not depend on whether we transform the data before analyzing them*.

**EXERCISE 4.8.** Let  $w \equiv G(y)$  with  $y$  a vector having density  $f(y|\theta)$  and  $G$  having a differentiable inverse function. Use Proposition B.4 to find the density of  $w$  and show that the likelihoods satisfy  $L(\theta|y) \propto L(\theta|w)$ .

#### 4.4 Sufficient Statistics

Sometimes a function of the data contains all the information about the parameter of interest. Suppose  $y|\theta \sim f(y|\theta)$ . A function of  $y$ , say,  $T(y)$  is said to be *sufficient* if the distribution of  $y$  given  $T(y)$  does not depend on  $\theta$ . Note that in general if  $T(r) = s$ ,

$$\Pr(y = r) = \Pr(y = r, T(y) = s) = \Pr(y = r | T(y) = s) \Pr(T(y) = s).$$

Similarly, for any  $T(y)$  we can write

$$f(y|\theta) = f_1[y|T(y), \theta] f_2[T(y)|\theta].$$

If  $T(y)$  is sufficient, the distribution of  $y$  given  $T(y)$  does not depend on  $\theta$ , so we have

$$f(y|\theta) = f_1[y|T(y)] f_2[T(y)|\theta].$$

There is no information about  $\theta$  in  $f_1[y|T(y)]$ , so we can ignore it. In other words, all of the information about  $\theta$  is contained in any sufficient statistic  $T(y)$ . When  $T(y)$  is sufficient for  $\theta$  we also have

$$L(\theta|y) = f_1[y|T(y)] f_2[T(y)|\theta]$$

in which  $f_1[y|T(y)]$  is merely a constant of proportionality, so for practical purposes the likelihood is  $f_2[T(y)|\theta]$  and only depends on the sufficient statistic. Moreover, since the Bayesian analysis of  $\theta$  depends only on the prior and the likelihood, it depends on the data only through the sufficient statistic.

A famous result in mathematical statistics is the *factorization criterion* which states that  $T(y)$  is sufficient if and only if one can write

$$f(y|\theta) = h(y)g(T(y); \theta)$$

for some functions  $h$  and  $g$ . Note that  $f_1[y|T(y)]$  is such a function  $h(y)$ . Another way to think of the factorization criterion is that  $T(y)$  is sufficient if and only if the likelihood  $L(\theta|y)$  is proportional to a function that depends on  $y$  only through  $T(y)$ , that is,  $L(\theta|y) \propto g(T(y); \theta)$  for some function  $g$ .

**EXAMPLE 4.4.1.** *Bernoulli Data.* In Example 2.3.2, we considered

$$y_1, \dots, y_n | \theta \stackrel{iid}{\sim} \text{Bern}(\theta).$$

The data have a sufficient statistic  $t = \sum_{i=1}^n y_i$  because, as seen in equation (2.3.2), the likelihood is proportional to a function that depends on the data only through  $t$ , so  $t$  must be sufficient.

**EXAMPLE 4.4.2.** *Normal Data.* Suppose we have independent normal observations with unknown mean  $\theta$  and known variance  $\sigma_0^2 \equiv 1/\tau_0$ , i.e.,

$$y_1, \dots, y_n | \theta \stackrel{iid}{\sim} N(\theta, 1/\tau_0).$$

The likelihood is

$$L(\theta | y) = \prod_{i=1}^n \left\{ \left( \frac{\tau_0}{2\pi} \right)^{1/2} \exp \left[ -\frac{\tau_0}{2} (y_i - \theta)^2 \right] \right\} \propto \exp \left[ -\frac{\tau_0}{2} \sum_{i=1}^n (y_i - \theta)^2 \right].$$

Writing  $\bar{y}_. = (y_1 + \dots + y_n)/n$  and performing some algebra we can write

$$\begin{aligned} L(\theta | y) &\propto \exp \left[ -\frac{\tau_0}{2} \sum_{i=1}^n (y_i - \bar{y}_. + \bar{y}_. - \theta)^2 \right] \\ &= \exp \left[ -\frac{\tau_0}{2} \sum_{i=1}^n (y_i - \bar{y}_.)^2 - \frac{\tau_0}{2} n (\bar{y}_. - \theta)^2 \right] \\ &\propto \exp \left[ -\frac{n\tau_0}{2} (\bar{y}_. - \theta)^2 \right]. \end{aligned}$$

The likelihood is proportional to a function that depends on the data only through  $\bar{y}_.$ , so  $\bar{y}_.$  must be sufficient.

**EXERCISE 4.9.** Prove that  $\sum_{i=1}^n (y_i - \theta)^2 = \sum_{i=1}^n (y_i - \bar{y}_.)^2 + n(\bar{y}_. - \theta)^2$ .

Although the Bayesian posterior analysis for  $\theta$  depends on the data only through the sufficient statistic, that fact presupposes that we know the correct sampling distribution  $f(y | \theta)$ . In Subsection 4.1.1 we used the marginal density  $f(y)$  as a tool for model checking.  $f(y)$  depends on the complete data and not merely a sufficient statistic. Model checking with  $f(y)$  incorporates both the sampling density and the prior into the model, so data that look unusual relative to  $f(y)$  call in question both the sampling density and the prior with no suggestion as to which may be inadequate. While we do not regard priors that have been elicited as “wrong,” they may be “misinformed.” *In practice all priors and sampling distributions are merely approximations and those approximations may be inadequate.*

Frequentists do not have a prior, so they only want to validate the sampling density  $f(y | \theta)$ . They often recommend using the distribution of  $y$  given  $T(y)$  as the basis for model checking because with  $T(y)$  sufficient, the distribution does not depend on the unknown  $\theta$ . We see no problem with that suggestion as it relates to the sampling distribution.

## 4.5 Analysis Using Predictive Distributions

Traditionally, the fundamental tool used in Bayesian analysis has been the posterior density  $p(\theta | y)$ . One obtains the posterior density from the sampling density  $f(y | \theta)$  and the prior density on the parameters  $p(\theta)$ .

One can argue on philosophical grounds that the fundamental tool of Bayesian analysis should be the predictive density  $f(\tilde{y} | y)$  for new observations  $\tilde{y}$ , see Geisser (1971, 1993). As discussed in Section 2.3, the standard way to get the predictive density is to integrate the sampling density of  $\tilde{y}$ ,  $f(\tilde{y} | \theta)$ , against the posterior of  $\theta$  (assuming that  $y$  and  $\tilde{y}$  are conditionally independent given  $\theta$ ). However, this approach does not always apply.

The Dirichlet process, developed by Ferguson (1973), is a tool that is used in “nonparametric” Bayesian analysis, cf. Chapter 15. It directly specifies a random probability distribution for the data  $y$ . Specifying a random probability distribution is also what the traditional Bayesian parametric methodology does, but it does it in two steps. The traditional methodology chooses a random distribution for the data by specifying a fixed parametric density  $f(y|\theta)$  conditional on the parameter  $\theta$  and then making  $\theta$  random by specifying a density  $p(\theta)$ .

In a Dirichlet process, the only conveniently available information is the marginal and predictive distributions. Prior to collecting data, the marginal is a distribution for what data  $y$  may be seen in the future, and after collecting data, the predictive is the distribution for what may be seen in the future  $\tilde{y}$ , given the current data  $y$ . The corresponding quantities in parametric Bayesian analysis are

$$f(y) = \int f(y|\theta)p(\theta)d\theta$$

and

$$f(\tilde{y}|y) = \int f(\tilde{y}|\theta)p(\theta|y)d\theta.$$

The successful application of Dirichlet processes demonstrates that it is possible to do Bayesian analysis using only predictive distributions, so predictive distributions must be more fundamental to the Bayesian process than posterior distributions. The use of parameters is a convenience, not a fundamental aspect of Bayesian analysis.

Geisser (1971) advocated “doing unto the predictive distribution what you would have done unto the original sampling distribution,” that is, using the predictive density as a surrogate for the sampling distribution. In a parametric setting, the idea is to estimate the unknown sampling density,  $f(\tilde{y}|\theta)$ , using the predictive density  $f(\tilde{y}|y)$ .

More formally, we can think of most interesting parametric functions  $\gamma(\theta)$  as functionals  $\gamma(\theta) = T[f(\tilde{y}|\theta)]$ . The suggestion is then to use  $T[f(\tilde{y}|y)]$  as an estimate of  $T[f(\tilde{y}|\theta)]$ .

Christensen and Huffman (1985) showed that for an interestingly broad class of functionals  $\gamma(\theta) = T[f(\tilde{y}|\theta)]$ , this predictive estimate is just the posterior mean of the parameter, that is,  $T[f(\tilde{y}|y)] = E(\gamma|y)$ . The class in question is the set of functionals that admit unbiased estimates, that is, functions  $\gamma(\theta) = E[h(\tilde{y})|\theta]$  for some function  $h$ . The corresponding predictive estimate is  $\delta(y) = E[h(\tilde{y})|y] = E[\gamma(\theta)|y]$ . A proof is given at the end of the section.

Geisser (1971) had previously noted this fact in various examples. If  $\tilde{y}$  is univariate, we might be interested in the mean of the sampling distribution  $\gamma(\theta) = E(\tilde{y}|\theta)$ . The equivalence result establishes that the posterior mean of the sampling expected value equals the mean of the predictive distribution, that is,

$$E[\gamma(\theta)|y] = \int \tilde{y}f(\tilde{y}|y)d\tilde{y}.$$

When analyzing time to event data, we are often interested in the survival function

$$S(t_0|\theta) \equiv \Pr(\tilde{y} > t_0|\theta) = \int I_{(t_0, \infty)}(\tilde{y})f(\tilde{y}|\theta)d\tilde{y}.$$

The predictive estimate is

$$S(t_0|y) \equiv \int I_{(t_0, \infty)}(\tilde{y})f(\tilde{y}|y)d\tilde{y} = E[\Pr(\tilde{y} > t_0|\theta)|y],$$

so the predictive survival function is just the posterior mean of the survival probability.

Unfortunately, life becomes more complicated if we let  $\tilde{y}$  be multidimensional. For example, we can have distinct functionals  $T_1$  and  $T_2$  with  $\gamma(\theta) = T_1[f(\tilde{y}|\theta)] = T_2[f(\tilde{y}|\theta)]$  for which replacing the sampling distribution with the predictive distribution can give different answers. For example, suppose the  $\tilde{y}_i$ s are iid given  $\theta$  and  $s_{\tilde{y}}^2$  is the sample variance of the  $\tilde{y}_i$ s. The variance  $\sigma^2 \equiv \sigma^2(\theta)$  of the sampling distribution can be obtained as either the variance of a single observation

$$\sigma^2 = \text{Var}(\tilde{y}_i|\theta) = \int [\tilde{y}_i - E(\tilde{y}_i|\theta)]^2 f(\tilde{y}|\theta)d\tilde{y} \equiv T_1[f(\tilde{y}|\theta)]$$

or as the expected value of the sample variance

$$\sigma^2 = \int s_{\tilde{y}}^2 f(\tilde{y}|\theta) d\tilde{y} \equiv T_2[f(\tilde{y}|\theta)].$$

$T_1$  is not in the class of unbiased functionals for which the equivalence holds because it is the expected value of a function that involves  $\theta$ . Also, the variance of the predictive distribution is typically larger than that of the sampling distribution, that is, typically

$$T_1[f(\tilde{y}|y)] = \text{Var}(\tilde{y}_i|y) > \text{Var}(\tilde{y}_i|\theta) = T_1[f(\tilde{y}|\theta)].$$

It follows that the variance of the predictive distribution makes a poor estimate of the variance of the sampling distribution. On the other hand, the predictive expectation of  $s_{\tilde{y}}^2$  makes a reasonable estimate because  $T_2$  has the property that  $E(\sigma^2|y) = T_2[f(\tilde{y}|y)]$ . See Christensen and Huffman (1985) for more details.

We now show that when

$$\gamma(\theta) \equiv T[f(\tilde{y}|\theta)] = \int h(\tilde{y}) f(\tilde{y}|\theta) d\theta = E[h(\tilde{y})|\theta]$$

for some  $h$ , then  $\delta(y) \equiv T[f(\tilde{y}|y)] = E[\gamma(\theta)|y]$ .

$$\begin{aligned} \delta(y) &\equiv T[f(\tilde{y}|y)] \\ &= \int h(\tilde{y}) f(\tilde{y}|y) d\tilde{y} \\ &= \int h(\tilde{y}) \left[ \int f(\tilde{y}|\theta) p(\theta|y) d\theta \right] d\tilde{y} \\ &= \int \left[ \int h(\tilde{y}) f(\tilde{y}|\theta) d\tilde{y} \right] p(\theta|y) d\theta \\ &= \int \gamma(\theta) p(\theta|y) d\theta \\ &= E[\gamma(\theta)|y]. \end{aligned}$$

**EXERCISE 4.10.** Let  $y_i|\theta \stackrel{iid}{\sim} \text{Exp}(\theta)$ ,  $i = 1, \dots, n+1$  and let  $p(\theta) = e^{-\theta}$ . Given  $y$ , obtain the predictive probability that  $y_{n+1} > t_0$  using calculus. Argue that this is also the posterior mean of a particular function of  $\theta$ . How would you interpret the difference between these two quantities despite the fact that the values are identical?

**EXERCISE 4.11.** Let  $y_i|\theta \stackrel{iid}{\sim} \text{Pois}(\theta)$ ,  $i = 1, \dots, n+1$  and let  $p(\theta) = e^{-\theta}$ . Given  $y$ , obtain the predictive probability that  $y_{n+1} = 0$  using calculus. Argue that this is also the posterior mean of a particular function of  $\theta$ .

## 4.6 Flat Priors

Consider  $y \sim \text{Bin}(n, \theta)$ . Often people think that putting a uniform distribution on  $\theta$  denotes ignorance of the value of  $\theta$ . However, as argued by Raiffa and Schlaifer (1961), if you are ignorant about  $\theta$  you should also be ignorant about  $\theta^2$ , and you cannot find a distribution that is uniform on both  $\theta$  and  $\theta^2$ .

Nonetheless, for a univariate parameter  $\theta$  taking values on the entire real line, people often use “uniform” priors, that is, use  $p(\theta) = 1$  as a prior. It really does not matter whether you take  $p(\theta) = 1$  or  $p(\theta) = 123,456$ , the point is that the prior is flat. The point is also that the prior is *improper* because  $\int p(\theta) d\theta = \infty$  regardless of what constant you choose for the density. In fact,

such flat priors are inherently stupid priors. Given any bounded set  $A$  and its complement  $A^c$ , the integral outside the set  $A$  is  $\int_{A^c} p(\theta) d\theta = \infty$  whereas the integral inside is  $\int_A p(\theta) d\theta < \infty$ , so the prior belief is that virtually all weight goes to parameter values that are bigger than any number you can think of. Flat priors are often approximated by a proper prior with a large variance. This is what caused problems in the Lindley-Jeffreys paradox of Subsection 4.1.3. Moreover, if you have a sampling distribution  $f(y|\theta)$  and use a flat prior  $p(\theta) = 1$  or another improper prior, the marginal density for the data  $f(y) = \int f(y|\theta) d\theta$  often does not exist.

The virtue of flat priors is that in many problems, the flat prior is easily overwhelmed by the data. For example, using  $p(\theta) = K$  for any old constant  $K$ , Bayes' Theorem gives us

$$\begin{aligned} p(\theta|y) &= \frac{L(\theta|y)p(\theta)}{\int L(\theta|y)p(\theta)d\theta} \\ &= \frac{L(\theta|y)K}{\int L(\theta|y)Kd\theta} \\ &= \frac{L(\theta|y)}{\int L(\theta|y)d\theta}, \end{aligned}$$

so the prior appears to play no role in the posterior. The posterior is simply a renormalization of the likelihood into a density for  $\theta$ . (Not all likelihoods can be renormalized, because not all have finite integrals with respect to  $\theta$ .) Using flat priors with the large sample approximations of Section 4.10 leads to Bayesian inferences that are similar to frequentist large sample maximum likelihood inferences. Box and Tiao (1973) devote considerable effort to using flat priors.

**EXAMPLE 4.6.1. *Normal Data.*** Following Example 2.3.3, suppose we have independent normal observations with unknown mean  $\theta$  and known variance  $\sigma_0^2$ :

$$y_1, \dots, y_n | \theta \stackrel{iid}{\sim} N(\theta, \sigma_0^2)$$

with precision  $\tau_0 = 1/\sigma_0^2$ . The likelihood is

$$L(\theta|y) \propto \exp \left[ -\frac{\tau_0}{2} \sum_{i=1}^n (y_i - \theta)^2 \right]$$

and, as established in Example 4.4.2, with  $\bar{y}_. = (y_1 + \dots + y_n)/n$  we can write

$$L(\theta|y) \propto \exp \left[ -\frac{n\tau_0}{2} (\bar{y}_. - \theta)^2 \right].$$

As a function of  $\theta$ , this is proportional to a  $N(\bar{y}_., 1/n\tau_0)$  density, so that must be the posterior under a flat prior. For example, a 95% posterior interval with a flat prior is  $\bar{y}_. \pm 1.96\sigma_0/\sqrt{n}$ , which is numerically equivalent to, but philosophically vastly different from, the traditional 95% confidence interval for this problem.

**EXAMPLE 4.6.2.** Consider an improper prior on the positive numbers

$$p(\theta) = \frac{1}{\theta} I_{(0,\infty)}(\theta).$$

It is improper because  $\infty = \int_0^\infty 1/\theta d\theta$ . We transform to  $\gamma = \log(\theta) = g(\theta)$  and “find” the density using the standard transformation technique given in Proposition B.4. Noting that  $g^{-1}(\gamma) = e^\gamma$  and  $dg^{-1}(\gamma)/d\gamma = e^\gamma$ , the density becomes

$$q(\gamma) = \frac{1}{e^\gamma} I_{(-\infty,\infty)}(\gamma) \times |e^\gamma| = 1.$$

Thus, our initial prior corresponds to a flat prior on  $\log(\theta)$ . Although the prior  $p(\theta) = (1/\theta)I_{(0,\infty)}(\theta)$  may work well in the sense that it is easily overwhelmed by the data, one should never forget that in itself it is saying very stupid things, namely, that  $\theta$  is likely to be either huge or essentially 0.

For

$$y_1, \dots, y_n | \mu, \tau \text{ iid } N(\mu, 1/\tau),$$

the traditional “noninformative” prior, which we refer to as the *standard improper reference (SIR)* prior, is

$$p(\mu, \tau) = \frac{1}{\tau}.$$

One can think of this as  $p(\mu, \tau) = p(\mu)p(\tau)$  with  $p(\mu) = 1$  and  $p(\tau) = 1/\tau$ . So in a sense we are taking independent flat priors on  $\mu$  and  $\log(\tau)$ .

**EXAMPLE 4.6.3.** Consider an improper prior on the unit interval

$$p(\theta) = \theta^{-1}(1-\theta)^{-1}I_{(0,1)}(\theta).$$

It is improper because  $\infty = \int_0^1 \theta^{-1}(1-\theta)^{-1}d\theta$ . We transform to  $\gamma = \log[\theta/(1-\theta)] = g(\theta)$  and “find” the density using Proposition B.4. Noting that  $g^{-1}(\gamma) = e^\gamma/(1+e^\gamma)$  and  $dg^{-1}(\gamma)/d\gamma = e^\gamma/[1+e^\gamma]^2$ , the density becomes

$$q(\gamma) = \left(\frac{e^\gamma}{1+e^\gamma}\right)^{-1} \left(\frac{1}{1+e^\gamma}\right)^{-1} I_{(-\infty, \infty)}(\gamma) \times |e^\gamma/[1+e^\gamma]^2| = 1.$$

Thus, our initial prior corresponds to a flat prior on  $\log[\theta/(1-\theta)]$ . Although the prior  $p(\theta) = \theta^{-1}(1-\theta)^{-1}I_{(0,1)}(\theta)$  may work well in the sense that it is easily overwhelmed by the data, one should never forget that in itself it is saying very stupid things, namely, that  $\theta$  is likely to be very near 0 or 1.

#### 4.6.1 Data Translated Likelihoods

Box and Tiao (1973) discuss a method for selecting “noninformative” priors that involves consideration of the relative amount of information in the data compared with the amount of information in the prior. We only discuss their method for a scalar parameter  $\theta$ . Their argument says that if the likelihood function  $L(\theta|y)$  is “data translated,” then it is sensible to use a “flat” prior for  $\theta$ .

With a data translated likelihood (DTL), if we plot the likelihood function for distinct data sets, the shape of the likelihood remains the same. Only the location is shifted by the data. For example, with  $N(\theta, 1/\tau_0)$  data, the likelihood is  $L(\theta|\bar{y}) \propto \exp[-n\tau_0(\theta - \bar{y})^2/2]$ . As we change  $\bar{y}$  and plot as a function of  $\theta$ , we get the same bell-shaped curve only the mode  $\bar{y}$  of the curve is moving.

Box and Tiao (1973) argued that for a DTL a constant prior  $p(\theta) = K$  results in relatively the same amount of information in the prior compared to that in the data, no matter what the data might be. The Box and Tiao prior for the mean of a normal population when the precision is known is just  $p(\theta) = K$ .

**EXAMPLE 4.6.4. Exponential Data.** The likelihood function for a single observation from an  $\text{Exp}(\theta)$  distribution is  $L(\theta|y) \propto \theta e^{-\theta y} \propto \theta y e^{-\theta y}$ . Reparameterize to  $\gamma = \log(\theta)$  to get

$$L(\gamma|y) \propto \exp[\gamma + \log(y) - e^{\gamma+\log(y)}],$$

which is a DTL. The Box-Tiao prior is  $p(\gamma) = K$  and using the usual transformation technique, the prior on  $\theta$  is

$$q(\theta) = p(\log(\theta)) \left| \frac{d \log(\theta)}{d \theta} \right| = K/\theta \propto 1/\theta.$$

We could also get a DTL if we had an iid sample from the exponential distribution.

**EXERCISE 4.12.** Plot  $L(\gamma|y)$  in Example 4.6.4 for  $y = 1$  and for  $y = 10$  on the same plot, and observe that it is a DTL. For the same values of  $y$ , plot  $L(\theta|y)$  and observe that this is not a DTL.

**EXERCISE 4.13.** Let  $y_1, \dots, y_n | \theta \stackrel{iid}{\sim} \text{Exp}(\theta)$ . Argue analytically that, with  $\gamma = \log(\theta)$ ,  $L(\gamma|y)$  is a DTL.

**EXERCISE 4.14.** Let  $y_1, \dots, y_n$  be a random sample from the  $N(0, 1/\tau)$  distribution. (a) Find a parameter  $\gamma$  that results in a DTL in that parameter. (b) Let  $n = 100$  and define  $T(y) = \sum_i y_i^2$ . Plot  $L[\tau|T(y)]$  and  $L[\gamma|T(y)]$  for  $T(y) = 100, 200, 300$  to evaluate how the shapes change or remain the same. (c) Derive the Box-Tiao prior for  $\tau$  based on having a “flat” prior for  $\gamma$ .

## 4.7 Jeffreys’ Priors

Jeffreys (1961) proposed a class of priors for Bayesian problems. Many people view them as being “noninformative.”

**EXAMPLE 4.7.1.** *Normal Data.* Suppose we have independent normal observations with known mean  $\mu_0$  and unknown variance  $\sigma^2$ :

$$y_1, \dots, y_n | \sigma^2 \stackrel{iid}{\sim} N(\mu_0, \sigma^2).$$

With precision  $\tau = 1/\sigma^2$ , up to a constant the likelihood for  $\tau$  is

$$L(\tau|y) \propto \prod_{i=1}^n \tau^{1/2} \exp\left[-\frac{\tau}{2}(y_i - \mu_0)^2\right] = \tau^{n/2} \exp\left[-\frac{\tau}{2} \sum_{i=1}^n (y_i - \mu_0)^2\right].$$

The log of the likelihood is

$$\log[L(\tau|y)] \propto \frac{n}{2} \log(\tau) + \left[-\frac{\tau}{2} \sum_{i=1}^n (y_i - \mu_0)^2\right].$$

The derivative of the log-likelihood is

$$\frac{d}{d\tau} \log[L(\tau|y)] = \frac{n}{2} \tau^{-1} + \left[-\frac{1}{2} \sum_{i=1}^n (y_i - \mu_0)^2\right].$$

The second derivative of the log-likelihood is

$$\frac{d^2}{d\tau^2} \log[L(\tau|y)] = -\frac{n}{2} \tau^{-2}.$$

Jeffreys’ prior for this problem is defined as

$$p(\tau) \propto \sqrt{\frac{n}{2} \tau^{-2}}$$

or, dropping the constant,

$$p(\tau) \propto \frac{1}{\tau}.$$

In general, for a one parameter problem, *Fisher's information* is defined to be the expected value of the negative of the second derivative of the log-likelihood. *Jeffreys' prior is defined as being proportional to the square root of the Fisher information.*

EXAMPLE 4.7.2. *Binomial Data.* Let  $y|\theta \sim \text{Bin}(n, \theta)$ , so

$$L(\theta|y) \propto \theta^y(1-\theta)^{n-y}.$$

The log-likelihood is  $y\log(\theta) + (n-y)\log(1-\theta)$ . The derivative is  $(y/\theta) - (n-y)/(1-\theta)$  and the negative of the second derivative is  $(y/\theta^2) + (n-y)/(1-\theta)^2$ . The expected value of this is

$$\frac{n\theta}{\theta^2} + \frac{n-n\theta}{(1-\theta)^2} = n\frac{1}{\theta} + n\frac{1}{(1-\theta)} = \frac{n}{\theta(1-\theta)}.$$

Jeffreys' prior is

$$p(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}, \quad (1)$$

which is a Beta(0.5, 0.5) distribution and gives greater plausibility to values near 0 and 1 than to values in between.

Interestingly, the use of Jeffreys' priors does not satisfy the stopping rule principle or the likelihood principle. The Jeffreys' priors for binomial and negative binomial data are different, so using Jeffreys' priors gives different answers for binomial and negative binomial data that have exactly the same number of successes and failures. Finding the negative of the second derivative of the log-likelihood is equivalent in the two cases, but Fisher's information involves taking an expected value and the expected values come out differently in the binomial and negative binomial cases. Using Jeffreys' priors, models with proportional likelihoods lead to different inferences (because they have different priors) so the likelihood principle is violated. Similarly, stopping rules can change statistical inferences so the stopping rule principle is violated.

EXERCISE 4.15. Find Jeffreys' prior for  $\theta$  based on a random sample of size  $n$  when (a)  $y_i|\theta \sim \text{Pois}(\theta)$ , (b)  $y_i|\theta \sim \text{Exp}(\theta)$ , (c)  $y_i|\theta \sim \text{Weib}(2, \theta)$ , (d)  $y_i|\theta$  is negative binomial as in Example 4.3.3.

#### 4.7.1 Multiple Parameter Jeffreys' Prior\*

Define

$$\ell(\theta) \equiv \log[L(\theta|y)].$$

Define the second derivative matrix  $\ddot{\ell}(\theta)$  as the matrix of second order partial derivatives of  $\ell(\theta)$ , that is, the symmetric matrix of values

$$\left\{ \frac{\partial^2}{\partial \theta_j \partial \theta_k} \ell(\theta) \right\}.$$

Fisher's information is defined as

$$I(\theta) \equiv E[-\ddot{\ell}(\theta)].$$

Jeffreys' prior is defined by

$$p(\theta) \propto \sqrt{\det[I(\theta)]}.$$

EXERCISE 4.16. The Jeffreys' prior for the  $N(\mu, 1/\tau)$  problem with both parameters unknown is different from the SIR prior,  $p(\mu)p(\tau) \propto 1/\tau$ , even though the SIR prior is the product of the Jeffreys' priors for  $\mu$  with known precision and  $\tau$  for known mean. Derive the Jeffreys' prior with both parameters unknown.

#### 4.8 Bayes Factors\*

Suppose we have two competing models for some data. Model  $M_0$  has a sampling density  $f(y|\theta_0, 0)$  and the competing model is  $M_1$  with sampling density  $f(y|\theta_1, 1)$ . There need be nothing in common about  $\theta_0$  and  $\theta_1$ , they need not even have the same dimension. For example,  $f(y|\theta_0, 0)$  might be the density for an  $\text{Exp}(\theta_0)$  and  $f(y|\theta_1, 1)$  might be the density of a log-normal with parameters  $(\theta_{11}, \theta_{12})$ . Recall that if  $\log(z) \sim N(\theta_{11}, \theta_{12})$ , then  $z$  has a log-normal distribution with the same parameters, written  $z \sim LN(\theta_{11}, \theta_{12})$ .

We have prior distributions on  $\theta_0$  and  $\theta_1$ :  $p_0(\theta_0)$  and  $p_1(\theta_1)$ , respectively. We also have prior probabilities on the two models, say,  $q_0$  and  $q_1 = 1 - q_0$ , respectively.

Consider the problem of testing  $M_0$  versus the alternative  $M_1$ . Technically, we define a new Bernoulli random variable  $M$  for the model taking on the values 0, 1. The joint density is a mixture involving indicator functions

$$p(y, \theta_0, \theta_1, M) = f(y|\theta_0, 0)p_0(\theta_0)q_0I_{\{0\}}(M) + f(y|\theta_1, 1)p_1(\theta_1)q_1I_{\{1\}}(M).$$

In a problem like this, it is hard to imagine having any interest in  $\theta_0$  or  $\theta_1$  unless you already know the model. In any case, our interest is in the models and the data, so compute

$$f(y|0) \equiv \int f(y|\theta_0, 0)p_0(\theta_0)d\theta_0$$

and

$$f(y|1) \equiv \int f(y|\theta_1, 1)p_1(\theta_1)d\theta_1.$$

As mentioned in Section 4.6, with an improper prior for  $p_i(\theta_i)$ , the corresponding density  $f(y|i)$  often does not exist and we cannot proceed. Now apply Bayes' Theorem to see that

$$\Pr(M = 0|y) = \frac{q_0 f(y|0)}{q_0 f(y|0) + q_1 f(y|1)}.$$

The posterior odds of model  $M_0$  are

$$\begin{aligned} \frac{\Pr(M = 0|y)}{\Pr(M = 1|y)} &= \frac{[q_0 f(y|0)]/[q_0 f(y|0) + q_1 f(y|1)]}{[q_1 f(y|1)]/[q_0 f(y|0) + q_1 f(y|1)]} \\ &= \frac{q_0}{q_1} \frac{f(y|0)}{f(y|1)} \\ &= \frac{q_0}{q_1} BF, \end{aligned}$$

where  $BF \equiv f(y|0)/f(y|1)$  is called the *Bayes Factor*. We have established that the posterior odds equal the prior odds  $q_0/q_1$  times the Bayes factor.

We often examine

$$LBF \equiv \log(BF)$$

rather than  $BF$  because it is more stable when  $BF$  is very large or very small. Positive values of LBF favor  $M_0$  and negative values favor  $M_1$ .

##### 4.8.1 General Parametric Testing

We can recast the discussion of Bayes factors into a more traditional context of testing parameters. Suppose we have a sampling density  $f(y|\theta)$  and a prior  $p(\theta)$ . Consider the problem of testing  $H_0 : \theta \in \Omega_0$  versus the alternative  $H_1 : \theta \in \Omega_1$  where  $\Omega_1$  is the complement of  $\Omega_0$  relative to the set of all allowable parameters  $\Omega$ , i.e.,  $\Omega_0 \cup \Omega_1 = \Omega$  and  $\Omega_0 \cap \Omega_1 = \emptyset$ . Now rewrite the prior. Let  $q_0 = \Pr(\theta \in \Omega_0)$  and  $q_1 = \Pr(\theta \in \Omega_1) = 1 - q_0$ . Also define  $p_i(\theta) \equiv p(\theta|\theta \in \Omega_i)$  for  $i = 0, 1$ , so

$$p(\theta) = q_0 p_0(\theta) I_{\Omega_0}(\theta) + q_1 p_1(\theta) I_{\Omega_1}(\theta).$$

More properly, rather than rewriting  $p(\theta)$  in this form, we assume that  $p(\theta)$  has this form. For example, if  $\Omega_0 = \{0\}$ , a continuous density on  $p(\theta)$  would give  $q_0 = 0$  and the testing problem becomes uninteresting. In any case, the posterior becomes

$$p(\theta|y) = \frac{f(y|\theta)p(\theta)}{f(y)} = \begin{cases} f(y|\theta)p_0(\theta)q_0/f(y) & \theta \in \Omega_0 \\ f(y|\theta)p_1(\theta)q_1/f(y) & \theta \in \Omega_1 \end{cases}.$$

The posterior odds of  $\theta$  being in  $\Omega_0$  are

$$\begin{aligned} \frac{\int_{\Omega_0} p(\theta|y)d\theta}{\int_{\Omega_1} p(\theta|y)d\theta} &= \frac{q_0 \int_{\Omega_0} f(y|\theta)p_0(\theta)/f(y)d\theta}{q_1 \int_{\Omega_1} f(y|\theta)p_1(\theta)/f(y)d\theta} \\ &= \frac{q_0 \int_{\Omega_0} f(y|\theta)p_0(\theta)d\theta}{q_1 \int_{\Omega_1} f(y|\theta)p_1(\theta)d\theta} \\ &= \frac{q_0}{q_1} \frac{f(y|\theta \in \Omega_0)}{f(y|\theta \in \Omega_1)} \\ &= \frac{q_0}{q_1} BF. \end{aligned}$$

#### 4.8.2 Nested Models

Consider again the problem of testing  $H_0 : \theta \in \Omega_0$  versus the alternative  $H_1 : \theta \in \Omega_1$  where  $\Omega_1$  is the complement of  $\Omega_0$ , but we now focus on problems in which  $\Omega_0$  is a set of probability 0 under any continuous distribution. As before, let  $q_0 = \Pr(\theta \in \Omega_0)$ ,  $q_1 = \Pr(\theta \in \Omega_1) = 1 - q_0$ , and  $p_i(\theta) \equiv p(\theta|\theta \in \Omega_i)$  for  $i = 0, 1$ , so again

$$p(\theta) = q_0 p_0(\theta) I_{\Omega_0}(\theta) + q_1 p_1(\theta) I_{\Omega_1}(\theta).$$

Although this discussion applies quite generally, to make it more concrete suppose the  $r$  vector  $\theta$  is partitioned into two pieces of lengths  $r_0$  and  $r_1$  so that  $\theta = (\theta'_0, \theta'_1)'$ . Further suppose that  $\Omega_0 = \{\theta | \theta_1 = 0\}$ , i.e., the null hypothesis is  $H_0 : \theta_1 = 0$ . In this case,  $q_0 = \Pr[\theta_1 = 0]$  is the prior probability of the null hypothesis but  $p_0(\theta)$  cannot be a continuous density on the  $r$  dimensional set  $\Omega_0$  because it would integrate to 0, rather than 1, over  $\Omega_0$ . Instead, we take  $p_0(\theta_0)$  as a continuous density on  $r_0$  dimensional space. However, we have no such problems with  $p_1$ . Although  $p_1$  is technically only defined on  $\Omega_1$ , we can actually define it on all of  $\Omega$  because  $\Omega_1$  is a set of probability 1.

The point of all this is that when  $\Omega_0$  is a set of probability 0 under any continuous distribution, we can test  $H_0 : \theta \in \Omega_0$  versus  $H_1 : \theta \in \Omega_1$  but it is equivalent to think about testing the reduced model  $\theta \in \Omega_0$  versus the larger (full) model  $\theta \in \Omega$ . Whichever way we think of it, the components we need to specify are the same. We need probabilities  $q_0$  and  $q_1$  for the hypotheses (or models), we need an  $r$  dimensional density  $p_1(\theta)$  for the alternative (full model), and we need a density  $p_0(\theta)$  restricted to  $\Omega_0$ , which, for example, when  $\Omega_0 = \{\theta | \theta_1 = 0\}$  reduces to an  $r_0$  dimensional density  $p_0(\theta_0)$ . Although technically  $q_1$  must be  $\Pr(\theta \in \Omega_1)$ , that is, the probability that the full model is true but the reduced model is not, we often refer to  $q_1$  as the probability of the full model.

#### 4.8.3 Simulating Bayes Factors

The integrations involved in finding Bayes factors can be daunting, but they are easy to approximate via sampling. Recall that we have two competing models:  $M_0$  with sampling density  $f(y|\theta_0, 0)$  and prior  $p_0(\theta_0)$  as well as  $M_1$  with sampling density  $f(y|\theta_1, 1)$  and prior  $p_1(\theta_1)$ . The prior probabilities on the two models were  $q_0$  and  $q_1 = 1 - q_0$ , respectively. The Bayes factor is  $f(y|0)/f(y|1)$ , where

$$f(y|i) \equiv \int f(y|\theta_i, i) p_i(\theta_i) d\theta_i.$$

It is this integral that can be difficult to compute, but that is simple to approximate. Sample  $\theta_i^k$ ,  $k = 1, \dots, s$  from the prior density  $p_i(\theta_i)$  and use the approximation

$$f(y|i) \doteq \frac{1}{s} \sum_{k=1}^s f(y|\theta_i^k, i).$$

Typically,  $f(y|\theta_i, i)$  is a product of  $n$  terms, so approximating the entire function  $f(y|i)$  using a sum of thousands of products may seem daunting. But Bayes factors are actually only evaluated for the one  $y$  that was observed, so the computation is manageable.

Most of our Bayesian inferences come about by taking one sample from the posterior distribution, or with two models a sample from the posterior of each. The method just given for computing Bayes factors is annoying because it requires us to take two additional samples; one from the prior associated with each model. In two special cases, Bayes factors can be computed from a single posterior sample.

If  $\theta_0$  and  $\theta_1$  happen to have the same dimension, a simpler computational scheme writes

$$\begin{aligned} BF_{01} &= \frac{f(y|0)}{f(y|1)} = \frac{\int f(y|\theta_0, 0)p_0(\theta_0)d\theta_0}{\int f(y|\theta_1, 1)p_1(\theta_1)d\theta_1} \\ &= \frac{\int f(y|\theta, 0)p_0(\theta)d\theta}{\int f(y|\theta, 1)p_1(\theta)d\theta} \\ &= \frac{\int \frac{f(y|\theta, 0)p_0(\theta)}{f(y|\theta, 1)p_1(\theta)} f(y|\theta, 1)p_1(\theta)d\theta}{\int f(y|\theta, 1)p_1(\theta)d\theta} \\ &= \int \left\{ \frac{f(y|\theta, 0)p_0(\theta)}{f(y|\theta, 1)p_1(\theta)} \right\} p_1(\theta|y)d\theta \\ &\equiv \int r(\theta)p_1(\theta|y)d\theta. \end{aligned} \quad (1)$$

Thus, we sample  $\theta^k$ ,  $k = 1, \dots, s$  from the posterior density  $p_1(\theta|y)$  of model  $M_1$  and use the approximation

$$BF_{01} \doteq \frac{1}{s} \sum_{k=1}^s r(\theta^k) = \frac{1}{s} \sum_{k=1}^s \frac{f(y|\theta^k, 0)p_0(\theta^k)}{f(y|\theta^k, 1)p_1(\theta^k)}.$$

In Section 8.3 this method will be used to compute Bayes factors to compare different models (link functions) for binomial regression.

Now suppose model  $M_0$  is nested in model  $M_1$ . Then let  $\theta_0$  denote the vector of parameters under  $M_0$ , and  $(\theta_0^*, \theta_1)$  denote the parameters under  $M_1$ . The parameters  $\theta_0$  correspond to the parameters  $\theta_0^*$  although they may have different meanings under the two models. With  $p_1(\theta_1)$  the marginal prior for  $\theta_1$  under  $M_1$ , we now obtain

$$\begin{aligned} BF_{01} &= \frac{f(y|0)}{f(y|1)} = \frac{\int f(y|\theta_0, 0)p_0(\theta_0)d\theta_0}{\int \int f(y|\theta_0^*, \theta_1, 1)p_1(\theta_0^*, \theta_1)d\theta_0^*d\theta_1} \\ &= \frac{\int f(y|\theta_0, 0)p_0(\theta_0)\{\int p_1(\theta_1)d\theta_1\}d\theta_0}{\int \int f(y|\theta_0, \theta_1, 1)p_1(\theta_0, \theta_1)d\theta_0d\theta_1} \\ &= \frac{\int \int \frac{f(y|\theta_0, 0)p_0(\theta_0)p_1(\theta_1)}{f(y|\theta_0, \theta_1, 1)p_1(\theta_0, \theta_1)} f(y|\theta_0, \theta_1, 1)p_1(\theta_0, \theta_1)d\theta_0d\theta_1}{\int \int f(y|\theta_0, \theta_1, 1)p_1(\theta_0, \theta_1)d\theta_0d\theta_1} \\ &= \int \left\{ \frac{f(y|\theta_0, 0)p_0(\theta_0)p_1(\theta_1)}{f(y|\theta_0, \theta_1, 1)p_1(\theta_0, \theta_1)} \right\} p_1(\theta_0, \theta_1|y)d\theta_0d\theta_1 \\ &\equiv \int r(\theta)p_1(\theta|y)d\theta. \end{aligned} \quad (2)$$

If the  $M_1$  prior has  $\theta_0^* \perp\!\!\!\perp \theta_1 | M_1$ , the integrand simplifies to

$$r(\theta) = \frac{f(y|\theta_0, 0)}{f(y|\theta_0, \theta_1, 1)} \frac{p_0(\theta_0)}{p_1(\theta_0)}.$$

**EXAMPLE 4.8.1.** *BF Comparing  $M_1 : \text{Gamma}(\alpha, \lambda)$  to  $M_0 : \text{Exp}(\lambda)$ .* We reparameterize the problem so that it is easier to specify priors. Define  $\mu = \alpha/\lambda$  to be the mean of the  $\text{Gamma}(\alpha, \lambda)$  distribution and take  $\mu | M_1 \sim \text{Gamma}(a, b)$ , and  $\lambda | M_1 \sim \text{Gamma}(c, d)$ , independently. We also define  $\mu_0 = 1/\lambda$  and let  $\mu_0 | M_0 \sim \text{Gamma}(a, b)$ . It is reasonable to think that the prior specification for the mean of the data would be the same regardless of the model. Now let  $\theta_0 = \mu_0$ ,  $\theta_0^* = \mu$ , and  $\theta_1 = \lambda$ . Then since we have  $p_0(\theta_0) = p_1(\theta_0)$ , we have from (2)

$$r(\theta) = \frac{\prod_i \theta_0^{-1} e^{-y_i/\theta_0}}{\prod_i \theta_1^{\theta_0\theta_1} y_i^{\theta_0\theta_1-1} e^{-\theta_1 y_i} / \Gamma(\theta_0\theta_1)} = \exp \left( \sum_{i=1}^n v_i \right),$$

where

$$v_i = -\{\log(\theta_0) + \theta_0\theta_1 \log(\theta_1)\} + y_i\{\theta_1 - 1/\theta_0\} - \{\theta_0\theta_1 - 1\} \log(y_i) + \log(\Gamma(\theta_0\theta_1)).$$

We use this form to simplify programming. The posterior mean of  $r(\theta)$  under  $M_1$  is  $BF_{01}$ .

Using a sample of 19 simulated  $\text{Exp}(1)$  observations with  $\text{Exp}(1)$  priors, WinBUGS code for obtaining  $BF_{01}$  is:

```
model{
  for(i in 1:n){
    y[i] ~ dgamma(alpha,lambda)
    v[i] <- -(log(theta0) + theta0*theta1*log(theta1)) +
      y[i]*(theta1 -1/theta0)-(theta0*theta1 -1)*log(y[i])+
      loggam(theta0*theta1) # WinBUGS log gamma function
  }
  alpha <- theta0*theta1
  lambda <- theta1
  theta0 ~ dgamma(a,b)
  theta1 ~ dgamma(c,d)
  BF <- exp(sum(v[1:n]))
}
list(n=19,c=1,d=1,a=1,b=1,
y=c(0.05129329, 0.10536052, 0.16251893, 0.22314355,
  0.28768207, 0.35667494, 0.43078292, 0.51082562,
  0.59783700, 0.69314718, 0.79850770, 0.91629073,
  1.04982212, 1.20397280, 1.38629436, 1.60943791,
  1.89711998, 2.30258509, 2.99573227))
```

We ran this code for 40,000 iterations and got the following results

node	mean	sd	MC error	2.5%	median	97.5%
BF	2.109	9.518	0.04857	0.7426	1.003	8.415

The Bayes factor is about 2, and thus favors the exponential model over the more general gamma model as we would expect.

**EXERCISE 4.17.** Consider comparing models for data  $y_1, \dots, y_n$ . (a) For  $M_1 : \text{Weib}(\alpha, \lambda)$  and  $M_0 : \text{Exp}(\lambda)$  show that the medians of these distributions are  $\theta_0^* = [\log(2)/\lambda]^\alpha$  and  $\theta_0 = \log(2)/\lambda$ , respectively. Let both medians have the same  $\text{Gamma}(a, b)$  prior. Let  $\theta_1 = \alpha$  have a  $\text{Gamma}(c, d)$

prior, independent of  $\theta_0^*$ . Obtain  $r(\theta)$  and  $v_i$ , and write WinBUGS code to obtain the Bayes factor using equation (2). With  $a = b = c = d = 1$ , run your code using the  $y$  vector of simulated  $\text{Exp}(1)$  data from the WinBUGS code of Example 4.8.1. (b) Using your output from (a) and the results for the same data from Example 4.8.1, calculate the Bayes factor to compare the  $\text{Weib}(\alpha, \lambda)$  model with a  $\text{Gamma}(\alpha, \lambda)$ . You need not write any more code to do this.

**EXERCISE 4.18\*.** Consider comparing models for data  $y_1, \dots, y_n$ . (a) The models  $M_1: LN(\mu, 1/\tau)$  and  $M_0: \text{Weib}(\alpha, \lambda)$  are not nested but have the same number of parameters. Show that the median of the log normal distribution is  $e^\mu$ . Place  $\text{Gamma}(1,1)$  priors on the medians of the distributions, say,  $\theta_0^* = e^\mu$  and  $\theta_0 = [\log(2)/\lambda]^{1/\alpha}$ , respectively. Place  $\text{Gamma}(1, 1)$  priors on  $\alpha$  and  $\tau$ ; all priors independent of one another. Use equation (1) to obtain a BF and write code to calculate it for the simulated  $\text{Exp}(1)$  data given in Example 4.8.1. You need the log-normal pdf, which can be derived using Proposition B.4. Which model is preferred and why? (b) Repeat part (a) replacing  $M_0$  with  $M_3: \text{Exp}(\lambda)$ . This requires combining the development of both equations (1) and (2) since the models are non-nested with different numbers of parameters.

**EXERCISE 4.19.** Simulate samples of size 10, 50, and 150 from a  $\text{Gamma}(2,2)$  distribution. Let all priors be independent  $\text{Gamma}(1,1)$ . (a) Obtain BFs for making all three comparisons among Gamma, Exponential, and Weibull models. Larger sample sizes should make it easier to identify the correct model. (b\*) Using the same data, calculate Bayes factors for comparing log-normal to Weibull, and log-normal to Exponential. Make as many BF comparisons as you can using these and the results from part (a).

**EXERCISE 4.20.** Consider two models that share some parameters, but where each model has a subset of parameters that are not shared by the other model. This occurs in non-nested regression models where several common variables may be included in the two models, but each has one or more predictor variables that are not shared by the other. Let  $\theta' = (\theta'_0, \theta'_1, \theta'_2)$  consist of three vectors of parameters and assume that  $M_1$  corresponds to parameters  $(\theta'_0, \theta'_1)$  while model  $M_2$  has parameters  $(\theta'_{0*}, \theta'_2)$ . Derive an expression analogous to (1) and (2) for obtaining  $BF_{12}$ .

#### 4.9 Other Model Selection Criteria

Kadane and Lazar (2004) review model selection from Bayesian and frequentist perspectives. Bayes factors are the most appropriate for Bayesians but are notoriously difficult to compute, can be quite sensitive to the prior specification, and are undefined for improper priors. Here we consider some alternatives. These measures can be used to compare alternative models for data with a specific sampling scheme and with the data measured on a common scale. (More on these requirements in Subsection 4.)

One very simple model selection criterion is a version of maximum likelihood. For any model  $M$  with parameters  $\theta_M$  and prior on the parameters  $p_M(\theta_M)$ , the marginal density of the data is

$$f(y|M) \equiv \int f(y|\theta_M, M)p_M(\theta_M)d\theta_M.$$

We can think of the model  $M$  as a parameter and  $f(y|M)$  as a likelihood for the models given the data  $y$ . A maximum likelihood model selection simply chooses the model  $\hat{M}$  that maximizes the chance (density) of seeing the data we actually observed, i.e., pick  $\hat{M}$  with

$$f(y|\hat{M}) = \max_M f(y|M).$$

### 4.9.1 Bayesian Information Criterion

Historically, the most important Bayesian model selection criterion has been the *Bayesian information criterion (BIC)*, proposed by Schwarz (1978). BIC can be thought of as providing a large sample approximation to the Bayes factor. We motivate the ideas through an example.

**EXAMPLE 4.9.1.** *Normal Data.* Suppose we have independent normal observations with unknown mean  $\theta$  and known precision  $\tau_*$ :

$$y_1, \dots, y_n | \theta \stackrel{iid}{\sim} N(\theta, 1/\tau_*).$$

As seen in Examples 2.3.3 and 4.4.2

$$L(\theta | y) \propto \exp \left[ -\frac{n\tau_*}{2} (\bar{y} - \theta)^2 \right].$$

Rather than observing the complete data, we could just as well assume that we observe the sufficient statistic

$$\bar{y}_* | \theta \sim N(\theta, 1/n\tau_*)$$

because it has a proportional likelihood function. We wish to test  $H_0 : \theta = 0$  versus  $H_1 : \theta \neq 0$ . The prior given  $H_0$  is  $p_0(\theta)$  and can only be a discrete density giving probability 1 to  $\theta = 0$ . Take  $p_1(\theta)$  as a  $N(0, 1/\tau_1)$  density. Strictly speaking,  $p_1(\theta)$  should not be defined for  $\theta = 0$ , but any integral over  $\theta = 0$  is zero so it does not matter.

The Bayes factor is  $f(\bar{y}_* | \theta = 0) / f(\bar{y}_* | \theta \neq 0)$ . Clearly, under this model  $f(\bar{y}_* | \theta = 0)$  is just a  $N(0, 1/n\tau_*)$  density. The other density involved is

$$f(\bar{y}_* | \theta \neq 0) = \int f(\bar{y}_* | \theta) p_1(\theta) d\theta.$$

Usually, this integration would be difficult but for this normal case there is a simple way to find the distribution. In general,  $p(y, \theta) = f(y|\theta)p(\theta)$ . If  $f(y|\theta)$  does not depend on  $\theta$ , that is if  $f(y|\theta) \equiv f(y)$ , we clearly have  $p(y, \theta) = f(y)p(\theta)$ , so  $y$  and  $\theta$  are independent. We now apply this result having everything conditional on  $\theta \neq 0$ , so we obtain conditional independence and instead of using  $y$  and  $\theta$ , we use  $y - \theta$  and  $\theta$ .

The model for this problem implies that

$$\bar{y}_* - \theta | \theta, \theta \neq 0 \sim N(0, 1/n\tau_*).$$

The distribution does not depend on  $\theta$ , so  $\bar{y}_* - \theta$  and  $\theta$  are independent given  $\theta \neq 0$ . The distribution we want is  $y = (y - \theta) + \theta$  given  $\theta \neq 0$  and this involves adding together two (conditionally) independent normal random variables. When adding two independent normals, the resulting mean and variance are just the sums of the component means and variances, so using

$$\theta | \theta \neq 0 \sim N(0, 1/\tau_1)$$

we see that

$$\bar{y}_* | \theta \neq 0 \sim N\left(0 + 0, \frac{1}{n\tau_*} + \frac{1}{\tau_1}\right)$$

which determines  $f(\bar{y}_* | \theta \neq 0)$ . For notational convenience, define

$$\frac{1}{\tilde{\tau}} \equiv \frac{1}{n\tau_*} + \frac{1}{\tau_1}.$$

Integration constants from the normal densities cancel, so the Bayes factor is

$$\frac{f(\bar{y}_* | \theta = 0)}{f(\bar{y}_* | \theta \neq 0)} = \frac{\sqrt{n\tau_*} \exp[-n\tau_*(\bar{y}_* - 0)^2/2]}{\sqrt{\tilde{\tau}} \exp[-\tilde{\tau}(\bar{y}_* - 0)^2/2]},$$

which can be simplified to

$$BF = \sqrt{1 + \frac{n\tau_*}{\tau_1}} \exp \left[ \frac{-\bar{y}^2}{2} (n\tau_* - \tilde{\tau}) \right] = \sqrt{1 + \frac{n\tau_*}{\tau_1}} \exp \left[ \frac{-n\tau_* \bar{y}^2}{2} \frac{n\tau_*}{n\tau_* + \tau_1} \right].$$

We would reject  $H_0 : \theta = 0$  if the Bayes factor gets too small, or equivalently, if  $-2 \log(BF)$  gets too large.

$$-2 \log(BF) = \frac{(n\tau_* \bar{y}^2)^2}{n\tau_* + \tau_1} - \log \left( 1 + \frac{n\tau_*}{\tau_1} \right).$$

Now suppose that  $n$  is large so that  $n\tau_*$  is much larger than  $\tau_1$ . One can show (using L'Hopital's rule) that

$$\log \left( 1 + \frac{n\tau_*}{\tau_1} \right) / \log(n) \doteq 1 \quad (1)$$

so

$$-2 \log(BF) \doteq n\tau_* \bar{y}^2 - \log(n).$$

Perhaps more familiarly, writing  $\sigma_*^2 = 1/\tau_*$ ,

$$-2 \log(BF) \doteq \left( \frac{\bar{y}^2 - 0}{\sigma_*^2 / \sqrt{n}} \right)^2 - \log(n),$$

which is the square of the usual frequentist test statistic minus  $\log(n)$ .

**EXERCISE 4.21.** Use L'Hopital's rule to establish (1).

More generally, we have models  $M_i$  with sampling density  $f(y|\theta_i, i)$  where  $\theta_i$  has dimension  $r_i$ . Let  $\hat{\theta}_i$  be the maximum likelihood estimate of  $\theta_i$  given model  $i$ . The main tool in the generalization is establishing an approximation to the marginal density of the data given the model:

$$f(y|i) \doteq f(y|\hat{\theta}_i, i) n^{-r_i/2}.$$

This result is by no means obvious and is established under some general conditions in Schwarz (1978). The *Bayesian information criterion* is defined as

$$BIC_i = 2 \log f(y|\hat{\theta}_i, i) - r_i \log(n).$$

This does not depend on the prior  $p_i(\theta_i)$ . The idea is based on having enough data to overwhelm the prior and the posterior is essentially concentrated at  $\theta_i = \hat{\theta}_i$ .

When considering two models, say,  $M_1$  and  $M_2$ , the log Bayes factor is

$$LBF_{12} = \log \frac{f(y|1)}{f(y|2)} \doteq \log \left[ \frac{f(y|\hat{\theta}_1, 1) n^{-r_1/2}}{f(y|\hat{\theta}_2, 2) n^{-r_2/2}} \right] = \log \left[ \frac{f(y|\hat{\theta}_1, 1)}{f(y|\hat{\theta}_2, 2)} \right] + \frac{1}{2}(r_2 - r_1) \log(n),$$

so

$$2LBF_{12} \doteq BIC_1 - BIC_2 \equiv \Delta BIC_{12}.$$

With nested models where model 1 is a special case of model 2, the alternative hypothesis is that model 2 holds but model 1 does not,

$$\Delta BIC_{21} = -2 \log \left( \frac{f(y|\hat{\theta}_1, 1)}{f(y|\hat{\theta}_2, 2)} \right) - (r_2 - r_1) \log(n)$$

where

$$-2 \log \left( \frac{f(y|\hat{\theta}_1, 1)}{f(y|\hat{\theta}_2, 2)} \right)$$

is the asymptotic version of the generalized likelihood ratio test statistic for the models. Model 2 will be chosen when the  $\Delta BIC_{21}$  statistic is large. An approximation to the  $\Delta BIC_{21}$  can be obtained by fitting both models and taking the (simulated) posterior means or medians for the  $\hat{\theta}_i$ s in the formulas.

There has been considerable sleight of hand in that the prior distributions that are needed to obtain marginal densities under the two models have completely disappeared by the end of the approximation. As such, frequentists often use the BIC and compare it with other model selection criteria.

#### 4.9.2 LPML

Our preferred criterion for model selection is often the *log pseudomarginal likelihood (LPML)* of Geisser and Eddy (1979). It derives from predictive considerations and leads to pseudo Bayes factors for choosing among models. This approach has seen increased popularity due in part to the relative ease with which LPML is stably estimated from MCMC output.

In a model  $M$ , suppose the data  $y_1, \dots, y_n$  arise independently given a model parameter  $\theta$  so that  $f(y|\theta, M) = \prod_{i=1}^n f_i(y_i|\theta, M)$ . To compute Bayes factors for various models  $M$  we need to compute  $f(y|M)$ . With improper priors this density may not exist. The idea is to replace  $f(y|M)$  with some predictive version of it. Specifically, Geisser and Eddy (1979) replace  $f(y|M)$  with the *pseudomarginal likelihood*

$$\hat{f}(y|M) \equiv \prod_{i=1}^n f_i(y_i|y_{-i}, M)$$

where  $f_i(y_i|y_{-i}, M)$  is the *i*th *conditional predictive ordinate (CPO<sub>i</sub>)*, that is, the predictive density based on all of the data except the *i*th observation, denoted  $y_{-i}$ , evaluated at the observed  $y_i$ . The LPML is given by

$$LPML = \sum_{i=1}^n \log(CPO_i).$$

Analogous to our discussion of model selection by maximum likelihood just prior to Subsection 4.9.1, it is reasonable to choose a model that maximizes the pseudomarginal likelihood or the LPML. To find a pseudo Bayes factor, exponentiate the difference between the LPML statistics for two competing models.

Gelfand and Dey (1994) show that  $CPO_i$  and thus LPML is easily estimated from a posterior sample  $\theta^1, \dots, \theta^s$  via

$$CPO_i^{-1} \doteq \frac{1}{s} \sum_{k=1}^s \frac{1}{f_i(y_i|\theta^k, M)}.$$

The result is based on showing that

$$CPO_i^{-1} = E \left[ \frac{1}{f_i(y_i|\theta, M)} \middle| y_1, \dots, y_n \right].$$

The rest of the subsection demonstrates this result. For simplicity, we suppress the  $M$ . The key fact is observing that

$$\begin{aligned} f_i(y_i|y_{-i}) &= \int f_i(y_i|\theta)p(\theta|y_{-i})d\theta \\ &= \int f_i(y_i|\theta) \frac{\prod_{j \neq i} f_j(y_j|\theta)p(\theta)}{f(y_{-i})} d\theta \\ &= \frac{\int \prod_{j=1}^n f_j(y_j|\theta)p(\theta)d\theta}{\int \prod_{j \neq i} f_j(y_j|\theta)p(\theta)d\theta}. \end{aligned}$$

It follows that

$$\begin{aligned}
 \frac{1}{f_i(y_i|y_{-i})} &= \frac{\int \prod_{j \neq i} f_j(y_j|\theta) p(\theta) d\theta}{\int \prod_{j=1}^n f_j(y_j|\theta) p(\theta) d\theta} \\
 &= \frac{\int \prod_{j \neq i} f_j(y_j|\theta) p(\theta) d\theta}{f(y)} \\
 &= \int \frac{1}{f_i(y_i|\theta)} \prod_{j=1}^n f_j(y_j|\theta) p(\theta) / f(y) d\theta \\
 &= \int \frac{1}{f_i(y_i|\theta)} p(\theta|y) d\theta.
 \end{aligned}$$

**EXERCISE 4.22.** Write WinBUGS code to obtain  $CPO_i$  for a single observation from a sample of size  $n$  taken from the  $\text{Exp}(\theta)$  distribution with a  $\text{Gamma}(a, b)$  prior on  $\theta$ . Modify the code so that you can obtain  $CPO_i$  for all  $i$  in one run. It may help to review the code given in Subsection 4.8.3.

#### 4.9.3 Deviance Information Criterion

Another model selection criterion is the *Deviance Information Criterion (DIC)* due to Spiegelhalter et al. (2002). Define a deviance as  $-2$  times the log-likelihood (corrected by any additive constant), that is,

$$D(\theta) \equiv -2 \log[L(\theta)] + C,$$

where  $C$  is any constant that does not depend on  $\theta$ . Compute

$$\mathbb{E}[D(\theta)|y].$$

This is used as a measure of model fit with small values being good. Also define an effective number of parameters

$$p_D \equiv \mathbb{E}[D(\theta)|y] - D[\hat{\theta}]$$

where  $\hat{\theta}$  is some Bayesian estimate of  $\theta$ , say the posterior mean, median, or mode. Finally, DIC is defined as the sum of the measure of model fit and a penalty for the effective number of parameters

$$DIC \equiv \mathbb{E}[D(\theta)|y] + p_D = 2\mathbb{E}[D(\theta)|y] - D[\hat{\theta}].$$

DIC is readily computed using a sample from the posterior distribution and is easily obtained in WinBUGS. In order to obtain DIC when you are running WinBUGS, click on the inference tool, then the DIC tool, and then click on “set.” After running the code, click on the “DIC” tool to obtain results.

A drawback of DIC is that the choice of  $\hat{\theta}$  can profoundly affect the statistic. Also, when applied to hierarchical models like those discussed in Section 12, it is unclear how many parameters there are, so it is unclear how to penalize for them. More on this in Section 13.

**EXAMPLE 4.9.1 CONTINUED.** *Normal Data.* Again with unknown mean  $\theta$  and known precision  $\tau_*$ :

$$y_1, \dots, y_n | \theta \stackrel{iid}{\sim} N(\theta, 1/\tau_*),$$

so

$$L(\theta|y) \propto \exp \left[ -\frac{n\tau_*}{2} (\bar{y} - \theta)^2 \right].$$

Take

$$D(\theta) = n\tau_* (\bar{y} - \theta)^2.$$

With prior  $\theta \sim N(\theta_0, 1/\tau_0)$ , the posterior is given in Table 2.3. With  $\hat{\theta} = E(\theta|y)$ , compute

$$\begin{aligned} E[D(\theta)|y] &= E[n\tau_*(\bar{y} - \theta)^2|y] \\ &= n\tau_* \text{Var}(\theta|y) + n\tau_*(\hat{\theta} - \bar{y})^2 \\ &= \frac{n\tau_*}{n\tau_* + \tau_0} + D(\hat{\theta}). \end{aligned}$$

It follows that

$$p_D = \frac{n\tau_*}{n\tau_* + \tau_0},$$

which for large  $n$  is close to 1. Note that

$$\begin{aligned} D(\hat{\theta}) &= n\tau_*(\bar{y} - \hat{\theta})^2 \\ &= n\tau_* \left( \bar{y} - \left[ \frac{n\tau_*}{\tau_0 + n\tau_*} \bar{y} + \frac{\tau_0}{\tau_0 + n\tau_*} \theta_0 \right] \right)^2 \\ &= n\tau_* \left( \frac{\tau_0}{n\tau_* + \tau_0} \right)^2 (\bar{y} - \theta_0)^2, \end{aligned}$$

so

$$DIC = n\tau_* \left( \frac{\tau_0}{n\tau_* + \tau_0} \right)^2 (\bar{y} - \theta_0)^2 + \frac{2n\tau_*}{n\tau_* + \tau_0}.$$

Unfortunately, having computed  $DIC$  for this simple model, it is not at all clear what other sampling models you would compare it to. Obviously, you can use it to compare different priors. To compare sampling models, they need to have comparable likelihood functions ( $y$  defined for a common scale) and the constant  $C$  in  $D(\theta)$  has to be identical.

#### 4.9.4 Final Comments

Alas, all of these model selection criteria violate the likelihood principle. (Which is not to say that we won't use them.) Consider log-normal data

$$y_1, \dots, y_n \stackrel{iid}{\sim} LN(\mu, \sigma^2), \quad \text{i.e.,} \quad \log(y_1), \dots, \log(y_n) \stackrel{iid}{\sim} N(\mu, \sigma^2).$$

Based on Exercise 4.8, the  $y_i$ s and the  $\log(y_i)$ s determine proportional likelihoods, but all of these model selection measures give different results if applied to the  $y_i$ s as opposed to the  $\log(y_i)$ s. The DIC depends on the constant  $C$  in the definition of  $D(\theta)$ . That constant is different for  $y_i \sim LN(\mu, \sigma^2)$  and  $\log(y_i) \sim N(\mu, \sigma^2)$ , so equivalent models get different DIC scores. Similarly, the BIC also involves different additive constants for equivalent data. In the case of LPML, it seems obvious that model comparisons have to be made for models with the data on the same scale.

*These model selection criteria are appropriate for comparing models that have a common measurement scale for the data  $y$  and the same sampling scheme (e.g., stopping rule). For example, they are not designed to compare, say, a binomial model to a negative binomial model. On the other hand, they are perfectly fine for comparing fixed effect logistic regression and probit models as discussed in Chapter 8, or the linear and nonlinear regression models discussed in Chapter 9.*

*One advantage of LPML is that it does not directly depend on the parameters of the model.* They have been integrated out. That becomes useful when, as discussed in Section 13, it becomes unclear what the parameters are.

EXERCISE 4.23. (a) The data

$$y = (1.287, 3.961, 0.538, 1.281, 0.482, 1.929, 1.951, 0.825, 0.801, 4.114)$$

were simulated from a  $LN(0,1)$  distribution in R using the command ‘`exp(rnorm(10,0,1))`’. Modify and run the following code and obtain the DIC statistic for analyzing the data as if it were normal, and for a log-normal analysis using `dflat()` as the prior on the mean/location and a  $\text{Gamma}(0.001, 0.001)$  on the precision in both cases. You may want to specify an initial value of 1 or something large for the precision, rather than letting WinBUGS generate the initial value. Also try `dflat()` for the log of the precision rather than the Gamma prior. The DIC statistic should be smaller for the correct, log-normal, model than for the incorrect normal model.

```
model{
  for(i in 1:10){ y[i] ~ dlnorm(mu,tau) }
  tau ~ dgamma(0.001,0.001)
  mu ~ dflat()
}
list(y=c(1.287, 3.961, 0.538, 1.281, 0.482,
       1.929, 1.951, 0.825, 0.801, 4.114))
list(mu=0,tau=1)
```

(b) Repeat part (a) using the Exponential data in Example 4.8.1. Also obtain the DIC when the data are correctly modeled as Exponential. (c) Explain why, although we are using both normal and log-normal distributions, these are valid DIC comparisons and not subject to the issues earlier.

#### 4.10 Normal Approximations to Posteriors\*

With large samples, the posterior distribution can be approximated by using a normal distribution. This section involves matrices of partial derivatives, multivariate normal distributions, and Fisher’s information.

Let

$$\ell(\theta) = \log[L(\theta|y)p(\theta)] = \log[p(\theta|y)] + \log[f(y)]$$

where  $\log[f(y)]$  is just a constant. Let  $\theta_*$  be the posterior mode, so that the partial derivatives of  $\ell(\theta)$  evaluated at  $\theta_*$  all equal 0. Taking a Taylor’s expansion of  $\ell(\theta)$  about  $\theta_*$  gives

$$\begin{aligned} \ell(\theta) &\doteq \ell(\theta_*) + \dot{\ell}(\theta_*)(\theta - \theta_*) + \frac{1}{2}(\theta - \theta_*)'\ddot{\ell}(\theta_*)(\theta - \theta_*) \\ &= \ell(\theta_*) + \frac{1}{2}(\theta - \theta_*)'\ddot{\ell}(\theta_*)(\theta - \theta_*), \end{aligned} \quad (1)$$

see Section A.10. Here  $\dot{\ell}(\theta)$  is a vector of partial derivatives and  $\ddot{\ell}(\theta)$  is a matrix of second order partial derivatives. It follows that

$$p(\theta|y)f(y) \doteq e^{\ell(\theta_*)} \exp\left[-\frac{1}{2}(\theta - \theta_*)'[-\ddot{\ell}(\theta_*)](\theta - \theta_*)\right]$$

and dropping terms that do not include  $\theta$  we get the approximation

$$p(\theta|y) \propto \exp\left[-\frac{1}{2}(\theta - \theta_*)'[-\ddot{\ell}(\theta_*)](\theta - \theta_*)\right]. \quad (2)$$

The right-hand side has the form of a multivariate normal density for  $\theta$  with mean vector  $\theta_*$  and covariance matrix  $[-\ddot{\ell}(\theta_*)]^{-1}$ . The reader may recall that  $-\ddot{\ell}(\theta_*)$  is analogous to the *Fisher Observed Information*. In particular, for a component of the vector  $\theta$  we can approximate  $\theta_j|y \sim N(\theta_{*j}, a_{jj})$  where  $a_{jj}$  is the  $(j, j)$  element of the inverse of the matrix  $-\ddot{\ell}(\theta_*)$ .

**EXAMPLE 4.10.1. Exponential Data.** We begin with exact results and then illustrate the normal approximation. Assume

$$y_1, \dots, y_n | \theta \stackrel{iid}{\sim} \text{Exp}(\theta)$$

so that the density of an individual  $y_i$  is

$$f_i(y_i|\theta) = \theta e^{-\theta y_i}$$

and

$$f(y|\theta) = L(\theta|y) = \prod_{i=1}^n f_i(y_i|\theta) = \theta^n e^{-\theta \sum y_i}.$$

Assume that the prior is  $\theta \sim \text{Gamma}(a, b)$  so the prior density is

$$p(\theta) \propto \theta^{a-1} e^{-b\theta}.$$

Note that  $E(\theta) = a/b$ , the mode of  $\theta$  is  $(a-1)/b$  for  $a \geq 1$  and 0 for  $0 < a < 1$ . The standard deviation is  $\sqrt{a/b^2} = \sqrt{E(\theta)/b}$ . The posterior is

$$\begin{aligned} p(\theta|y) &\propto L(\theta|y)p(\theta) \\ &\propto \left[ \theta^n e^{-\theta \sum y_i} \right] \left[ \theta^{a-1} e^{-b\theta} \right] \\ &= \theta^{a+n-1} e^{-(b+\sum y_i)\theta} \end{aligned}$$

which implies that

$$\theta|y \sim \text{Gamma}(a+n, b + \sum y_i).$$

The posterior mean of  $\theta$  is  $E(\theta|y) = (a+n)/(b+\sum y_i)$ , the posterior mode is  $(a+n-1)/(b+\sum y_i)$  for  $n \geq 1$ . The standard deviation is  $\sqrt{(a+n)/(b+\sum y_i)^2}$ . If  $a+n$  is an integer, we can easily find interval estimates because

$$2(b + \sum y_i) \theta | y \sim \text{Gamma}\left(\frac{2(a+n)}{2}, \frac{1}{2}\right) = \chi^2_{2(a+n)},$$

see Exercise 5.19. Letting  $\chi_v^2(\alpha)$  denote the  $\alpha$  percentile of a  $\chi^2_v$  distribution,

$$\begin{aligned} 1 - \alpha &= \Pr \left[ \chi^2_{2(a+n)} \left( \frac{\alpha}{2} \right) < 2(b + \sum y_i) \theta < \chi^2_{2(a+n)} \left( 1 - \frac{\alpha}{2} \right) \middle| y \right] \\ &= \Pr \left[ \chi^2_{2(a+n)} \left( \frac{\alpha}{2} \right) \frac{1}{2(b + \sum y_i)} < \theta < \chi^2_{2(a+n)} \left( 1 - \frac{\alpha}{2} \right) \frac{1}{2(b + \sum y_i)} \middle| y \right]. \end{aligned}$$

To apply the normal approximation, observe that

$$\ell(\theta) = (a+n-1) \log(\theta) - \theta(b + \sum y_i).$$

This implies that

$$\dot{\ell}(\theta) = \frac{a+n-1}{\theta} - (b + \sum y_i).$$

To find the posterior mode, set the derivative to 0 giving

$$\theta_* = \frac{a+n-1}{b + \sum y_i},$$

as advertised earlier. Moreover,

$$\ddot{\ell}(\theta) = -\frac{a+n-1}{\theta^2},$$

so the analogue to the inverse of the observed Fisher information is

$$[-\ddot{\ell}(\theta_*)]^{-1} = \frac{\theta_*^2}{a+n-1}$$

and the normal approximation is

$$\theta|y \sim N\left(\theta_*, \frac{\theta_*^2}{a+n-1}\right).$$

It follows that an approximate, say, 95% Bayesian posterior interval has endpoints

$$\theta_* \pm 1.96 \frac{\theta_*}{\sqrt{a+n-1}}.$$

With  $a = 2$ ,  $b = 1$ ,  $n = 20$ , and  $\sum y_i = 25$ , an exact equal tailed 95% interval is

$$\chi_{2(2+20)}^2(0.025) \frac{1}{2(1+25)} < \theta < \chi_{2(2+20)}^2(0.975) \frac{1}{2(1+25)}$$

or (0.5303, 1.2346) whereas the large sample interval is

$$\frac{2+20-1}{1+25} \pm 1.96 \frac{21/26}{\sqrt{2+20-1}}$$

or (0.4622, 1.1532). With more data, the approximation typically gets better. For  $a = 2$ ,  $b = 1$ ,  $n = 80$ , and  $\sum y_i = 100$ , the exact 95% interval is (0.6457, 0.9968) while the normal approximation gives (0.6273, 0.9767). Figure 4.2 compares the two gamma distributions with their normal approximations.

Now suppose we were interested in the approximate posterior for the scalar function  $\gamma \equiv g(\theta)$ . The *delta method* asserts that

$$\gamma|y \sim N(g(\theta_*), [\dot{g}(\theta_*)]'[-\ddot{g}(\theta_*)]^{-1}[\dot{g}(\theta_*)]),$$

where  $\dot{g}(\theta)$  is the column vector of partial derivatives of  $g(\theta)$  with respect to the individual components of the vector  $\theta$ . For example, if  $\gamma = g(\theta) = \theta_1/\theta_2$ , and the sample size is large, we have

$$\gamma|y \sim N\left(\frac{\theta_{*1}}{\theta_{*2}}, \left[\frac{1}{\theta_{*2}}, -\frac{\theta_{*1}}{\theta_{*2}^2}\right] \left[-\ddot{g}(\theta_*)\right]^{-1} \left[\frac{1}{\theta_{*2}}, -\frac{\theta_{*1}}{\theta_{*2}^2}\right]'\right).$$

**EXERCISE 4.24.** Let  $y|\theta \sim \text{Bin}(n, \theta)$  and assume that  $n$  is large. (a) Obtain the large sample approximation to the posterior assuming  $\theta \sim U[0, 1]$ . Then, just for fun, (b) find the large sample approximation to the posterior for  $\gamma \equiv \theta^2$ . This parameter is of little interest but the delta method applies easily.

**EXERCISE 4.25.** Let  $y_1, \dots, y_n$  be a random sample from the  $N(\mu, 1/\tau)$  distribution. (a) Find the large sample normal approximation to the posterior for  $(\mu, \tau)$  under the assumption that  $p(\mu, \tau) \propto 1/\tau$ . You will have to obtain the  $2 \times 2$  Fisher Information matrix analogue and its inverse. (b) Find the explicit (which means do the algebra) large sample normal approximation to the posterior for the coefficient of variation,  $\gamma \equiv 1/\mu\sqrt{\tau}$ .

We now take an excursion back to Subsection 4.6.1 where we discussed data translated likelihoods. We develop a method of obtaining approximate DTLs for a scalar  $\theta$ . We can expand the log likelihood in exactly the same way that we expanded the log posterior in (1). The left-hand side of (1) becomes  $\log[L(\theta|y)]$ ,  $\theta_*$  is the maximum likelihood estimate, and  $-\ddot{g}(\theta_*)$  is the actual Fisher Observed Information, which is minus the second derivative of the log likelihood evaluated at the maximum likelihood estimate. We thus obtain the approximate result for large  $n$

$$L(\theta|y) \propto \exp(-0.5(\theta_* - \theta)^2 / [\ddot{g}(\theta_*)]^{-1}).$$

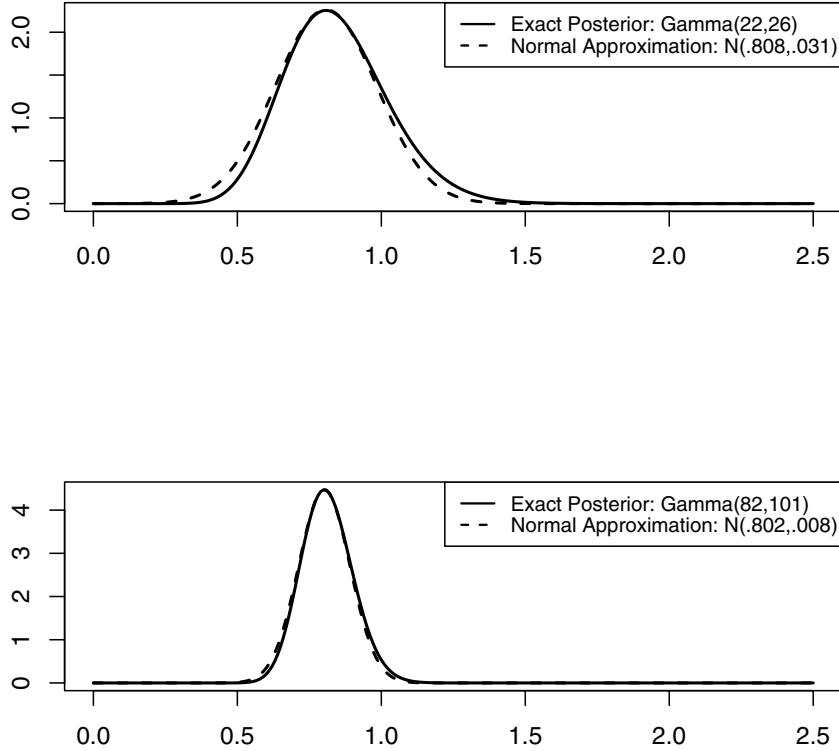


Figure 4.2: Comparisons of exact gamma distributions with normal approximations.

Consider a re-parametrization to  $\gamma = g(\theta)$  where  $g$  is a one-to-one function so that we can obtain  $\theta = g^{-1}(\gamma)$ . Then the likelihood function for  $\gamma$  can be written as  $\bar{L}(\gamma|y) = L(g^{-1}(\gamma)|y)$ . Expanding  $\log(\bar{L}(\gamma|y))$  about  $\gamma_* = g(\theta_*)$ , we obtain the approximation

$$\bar{L}(\gamma|y) \propto \exp\left(-0.5(g(\theta_*) - \gamma)^2 / \{[\dot{g}(\theta_*)]^2 / [-\ddot{g}(\theta_*)]\}\right). \quad (3)$$

Now suppose we can find a function  $\gamma = g(\theta)$ , such that  $[\dot{g}(\theta)]^2 / [-\ddot{g}(\theta)] = K$  a constant, i.e.,  $|\dot{g}(\theta)| \propto \sqrt{-\ddot{g}(\theta)}$ . Clearly the approximate likelihood function in  $\gamma$  is approximately data translated. In this case, we can place a constant prior on  $\gamma$  and be relatively noninformative in the sense described in Subsection 4.6.1. But if  $p(\gamma) \propto 1$ , say, then using Proposition B.4 the induced prior on  $\theta$  is  $p(\theta) \propto |\dot{g}(\theta)| \propto \sqrt{-\ddot{g}(\theta)}$ . But for large samples  $-\ddot{g}(\theta)/n \doteq E[-\ddot{g}(\theta)]/n$ , the Fisher Expected Information for a single observation. This approximation is justified because under iid sampling, the observed information can be seen as a sum of iid variates, so by the law of large numbers, it is approximately equal to the expected information of an individual, i.e., the expected information of the sample divided by  $n$ . It follows that the induced prior is approximately proportional to the square root of the Fisher Expected Information, which is Jeffreys prior.

**EXERCISE 4.26.** Let  $y_1, \dots, y_n$  be an iid sample from  $\text{Pois}(\theta)$ . Find the transformation  $\gamma = g(\theta)$  that makes the Poisson likelihood approximately data translated. Then find the induced prior on  $\theta$  corresponding to  $p(\gamma) \propto 1$ . Argue that this is approximately the Jeffreys prior for this problem.

### 4.11 Bayesian Consistency and Inconsistency

Bayesian point estimates tend to have good large sample properties. An estimate  $\hat{\theta}$  is consistent for  $\theta$  if  $\hat{\theta} \xrightarrow{P} \theta$ , cf. Section 3.4. The probabilities involved are computed given the parameter value, a very non-Bayesian attitude.

**EXAMPLE 4.11.1.** *Binomial Data.* We follow the notation and results of Example 2.3.1. Suppose  $y|\theta \sim \text{Bin}(n, \theta)$  with prior  $\theta \sim \text{Beta}(a, b)$ . The posterior is

$$\theta|y \sim \text{Beta}(y+a, n-y+b).$$

The prior is selected to reflect the researcher's beliefs and uncertainty. In particular, we saw that the prior mean and variance are

$$E(\theta) = \frac{a}{a+b} \equiv \mu \quad \text{and} \quad \text{Var}(\theta) = \frac{\mu(1-\mu)}{\psi+1},$$

where  $\psi \equiv a+b$  operates as a prior sample size measuring the strength of the researcher's beliefs. This interpretation is at least partly by analogy with the behavior of the posterior. The posterior is also a beta distribution with mean

$$\hat{\theta} \equiv E(\theta|y) = \left( \frac{n}{\psi+n} \right) \left( \frac{y}{n} \right) + \left( \frac{\psi}{\psi+n} \right) \mu$$

and variance

$$\text{Var}(\theta) = \frac{\hat{\theta}(1-\hat{\theta})}{n+\psi+1}.$$

The posterior mean is a weighted average of the prior mean  $\mu$  and the MLE of  $\theta$ ,  $y/n$ . For fixed values of  $\mu$  and  $\psi$  ( $a$  and  $b$ ) and large values of  $n$ , the posterior variance gets small so the posterior becomes highly concentrated near the posterior mean, which is approximately  $y/n$ , so the posterior mean shares the same consistency properties as  $y/n$ . (Similarly, for large values of  $\psi$ , the prior becomes highly concentrated about the prior mean.)

**EXAMPLE 4.11.2.** *Normal Data.* Consider  $y_1, \dots, y_n$  a random sample from a normal with mean  $\theta$  and variance 1. For the prior density, we take  $p(\theta)$  from a  $N(\mu_0, 1)$ . As a special case of Example 2.3.3, the posterior distribution of  $\theta$  given the data is

$$\theta|y_1, \dots, y_n \sim N\left(\frac{1}{n+1}\mu_0 + \frac{n}{n+1}\bar{y}_., \frac{1}{n+1}\right).$$

As  $n$  gets large, the variance gets small so the posterior becomes concentrated about the posterior mean  $(\mu_0 + n\bar{y}_.)/(n+1)$ , which, for large  $n$ , behaves like  $\bar{y}_.$ , so the posterior mean shares the same consistency properties as  $\bar{y}_.$

We now provide a particularly simple example of an inconsistent Bayes estimate and draw some conclusions from that example. In particular, the example has a posterior mean that is inconsistent on a dense subset of the real line.

**EXAMPLE 4.11.3.** *Inconsistent Posterior Mean.* Consider  $y_1, \dots, y_n$  an iid sample from a density  $f_*(y_i|\theta)$ . The distribution associated with  $f_*(y_i|\theta)$  is Cauchy with median  $\theta$  [i.e.,  $t(1, \theta, 1)$ ] when  $\theta$  is a rational number and Normal with mean  $\theta$  and variance 1 when  $\theta$  is irrational. In other words,

$$y_1, \dots, y_n|\theta \stackrel{iid}{\sim} \begin{cases} \text{Cauchy}(\theta) & \text{if } \theta \text{ is rational} \\ N(\theta, 1) & \text{if } \theta \text{ is irrational} \end{cases}.$$

For the prior density, we take  $p(\theta)$  to be absolutely continuous. For the sake of simplicity, take it to be  $N(\mu_0, 1)$ .

It can be shown that the posterior distribution of  $\theta$  given the data is the same as if the entire sampling distribution of  $y|\theta$  were  $N(\theta, 1)$ . In other words, the posterior distribution is

$$\theta|y_1, \dots, y_n \sim N\left(\frac{1}{n+1}\mu_0 + \frac{n}{n+1}\bar{y}, \frac{1}{n+1}\right).$$

This occurs because there are so few rational numbers compared to how many irrationals there are. (The rationals are a set of Lebesgue measure 0.) A proof that this is the posterior distribution is given in Christensen (2009). The posterior mean,  $(\mu_0 + n\bar{y})/(n+1)$  behaves asymptotically like  $\bar{y}$ . If the true value of  $\theta$  is an irrational number, the true sampling distribution is normal and the Bayes estimate is consistent just as in Example 4.11.2. However, if the true value of  $\theta$  is a rational number, the true sampling distribution is Cauchy( $\theta$ ), for which it is well known that  $\bar{y}$  is an inconsistent estimate of  $\theta$  because the distribution of  $\bar{y}$  is still Cauchy( $\theta$ ) and thus for large  $n$  the distribution is NOT becoming tightly concentrated around  $\theta$ . Therefore, we have a Bayes estimate that is inconsistent on a dense set (the rationals), but a dense set that has prior probability zero.

It seems quite clear from the calculus behind this example that the proper concern for Bayesians is whether their procedures are consistent with prior probability one. In general, there seems to be little hope of forcing consistency on sets of prior probability zero, although considerable Bayesian literature is concerned with establishing conditions under which one can get consistency for every parameter value.

## 4.12 Hierarchical Models

*Hierarchical models* involve specifying either the prior distribution or the sampling distribution as a series of models. They are sometimes called “*multilevel models*.” Whether they generalize the prior or the sampling distribution is actually a matter of interpretation. We begin by looking at examples.

**EXAMPLE 4.12.1.** Suppose we take a sample of  $m$  hospitals from around the country and then sample  $n_i$  patients within each hospital, asking each patient whether or not they were satisfied with their treatment. Within each hospital, results should be independent with a constant probability of satisfaction  $\theta_i$  that depends on the hospital  $i$ . The total number of satisfied patients from hospital  $i$  in our sample is

$$y_i|\theta_i \sim \text{Bin}(n_i, \theta_i),$$

independently for  $i = 1, \dots, m$ . Conditioning on the hospital, individual patient observations should be independent but, unconditionally, observations within a hospital should be more alike than observations in different hospitals. That occurs because the  $\theta_i$ s vary.

The hospitals are a random sample of hospitals, so there should be a distribution for the hospital satisfaction proportions  $\theta_i$ . One reasonable model for the  $\theta_i$ s is to take them as iid from a beta distribution, say,

$$\theta_i|\alpha_1, \alpha_2 \stackrel{iid}{\sim} \text{Beta}(\alpha_1, \alpha_2).$$

Together, the binomial distribution and the beta distribution define the sampling distribution for the independent  $y_i$ s. These are beta-binomial distributions with parameters  $n_i$ ,  $\alpha_1$ , and  $\alpha_2$  as discussed in Example 4.1.2. Frankly, beta-binomial distributions are not the easiest distributions to work with, and we will see that in other similar problems the sampling distribution cannot be identified as any convenient parametric family.

To put a prior on the parameters  $\alpha_1$ , and  $\alpha_2$ , we induce it from information on parameters that are easier to think about. The mean of the beta distribution is  $\mu \equiv \alpha_1/(\alpha_1 + \alpha_2)$  with  $\psi = \alpha_1 + \alpha_2$  a measure of spread for the beta distribution. Since  $\text{Var}(\theta_i|\mu, \psi) = \mu(1 - \mu)/(\psi + 1)$ , smaller  $\psi$ s

give larger variances to the  $\theta_i$ s. We are uncertain about  $(\mu, \psi)$ , so we elicit a prior distribution for them. A reasonable prior might be

$$\mu \sim \text{Beta}(a, b) \quad \perp\!\!\!\perp \quad \psi \sim \text{Gamma}(c, d).$$

Details of prior elicitation are given in Sections 5.1 and 14.4. Briefly, experts can think about their best guess for  $\mu$  and use a 95th prior percentile for  $\mu$  to determine the appropriate Beta prior. They can also think about the 90th percentile of the  $\theta_i$  distribution to obtain a best guess for  $\psi$ .

**EXAMPLE 4.12.2.** We now consider an alternative sampling distribution for the problem in Example 4.12.1. Again we assume

$$y_i | \theta_i \stackrel{iid}{\sim} \text{Bin}(n_i, \theta_i),$$

for each hospital  $i = 1, \dots, m$ , but now we assume a logit-normal model for the  $\theta_i$ s, e.g.,

$$\text{logit}(\theta_i) | \alpha_1, \alpha_2 \stackrel{iid}{\sim} N(\alpha_1, 1/\alpha_2).$$

This provides an alternative sampling distribution that has no name and is, again, not easy to work with directly.

To specify our prior uncertainty about the unknown parameters  $\alpha_1$  and  $\alpha_2$ , a convenient joint prior is

$$\alpha_1 \sim N(a, 1/b) \quad \perp\!\!\!\perp \quad \alpha_2 \sim \text{Gamma}(c, d).$$

A best guess for the median of the  $\theta_i$  distribution determines  $a$  and, together with a 95th prior percentile for the median, we can determine  $b$ . Information on percentiles of the  $\theta_i$  distribution can provide values for  $c$  and  $d$ .

**EXAMPLE 4.12.3.** Suppose, instead of a blunt “yes/no” on the question of satisfaction, we measure some continuous response from each patient  $j$  in each hospital  $i$ . A simple random effects model for such data includes an overall mean  $\mu$  for all observations, a random hospital effect  $\eta_i$ , and individual random errors  $\varepsilon_{ij}$ . Formally, write

$$y_{ij} = \mu + \eta_i + \varepsilon_{ij}, \quad \eta_i \stackrel{iid}{\sim} N(0, 1/\alpha_1), \quad \varepsilon_{ij} \stackrel{iid}{\sim} N(0, 1/\alpha_2),$$

$i = 1, \dots, m$ ,  $j = 1, \dots, n_i$  with the  $\eta_i$ s and  $\varepsilon_{ij}$ s all independent. With normal distributions we can identify the sampling distribution as a multivariate normal:

$$y_{ij} \sim N\left(\mu, \frac{1}{\alpha_1} + \frac{1}{\alpha_2}\right), \quad \text{Cov}(y_{ij}, y_{i'j'}) = \begin{cases} 0 & \text{if } i \neq i' \\ 1/\alpha_1 & \text{if } i = i', j \neq j' \end{cases}.$$

Although these are traditional ways of writing the sampling model, Bayesians might use different notation:

$$y_{ij} | \theta_i, \alpha_2 \stackrel{ind}{\sim} N(\theta_i, 1/\alpha_2) \tag{1}$$

with

$$\theta_i | \mu, \alpha_1 \stackrel{iid}{\sim} N(\mu, 1/\alpha_1). \tag{2}$$

In (1), no relationship has been specified between the observable  $y_{ij}$  and the unobservable pair  $\mu$  and  $\alpha_1$ , so as in Section B.3 it is implicitly assumed that  $y_{ij} \perp\!\!\!\perp \mu, \alpha_1 | \theta_i, \alpha_2$ . Similarly, from (2), the unobservables are implicitly assumed to satisfy  $\theta_i \perp\!\!\!\perp \alpha_2 | \mu, \alpha_1$ . It follows that

$$y_{ij} | \theta_i, \mu, \alpha_1, \alpha_2 \stackrel{ind}{\sim} N(\theta_i, 1/\alpha_2), \quad \theta_i | \mu, \alpha_1, \alpha_2 \stackrel{iid}{\sim} N(\mu, 1/\alpha_1).$$

As such, we need only to specify a prior on the parameters  $\mu$ ,  $\alpha_1$  and  $\alpha_0$ . A convenient choice is independent normal, gamma, and gamma priors, respectively.

So far, we have dealt with hierarchical models to provide more flexible sampling distributions. Sometimes they are used to model uncertainty in a prior distribution.

EXAMPLE 4.12.4. Suppose

$$y|\theta \sim \text{Bin}(n, \theta)$$

with

$$\theta \sim \text{Beta}(\alpha_1, \alpha_2).$$

To have a well-defined prior,  $\alpha_1$  and  $\alpha_2$  must be known. Sometimes, people are either unwilling to specify these values, or equivalently, want a more general family of prior distributions for  $\theta$ . In either case, one can specify priors on  $\alpha_1$  and  $\alpha_2$ . These can be independent gamma priors or they could be reference flat priors on  $\log(\alpha_j)$ . If one is unwilling to specify  $\alpha_1$  and  $\alpha_2$  in a beta distribution, it is hard to imagine that one would be willing to specify the parameters of the gamma distributions on  $\alpha_1$  and  $\alpha_2$ , so reference priors are typically used.

We have been writing a standard Bayesian model as

$$y|\theta \sim f(y|\theta), \quad \theta \sim p(\theta).$$

Typically,  $p(\theta)$  is a member of a parametric family so it depends on some numerical constants that are chosen to reflect a researcher's beliefs about  $\theta$ . More accurately, a standard Bayesian model is

$$y|\theta \sim f(y|\theta), \quad \theta \sim p_0(\theta|\alpha_0),$$

where  $p_0$  indicates a parametric family of distributions, e.g., the beta family for binomial data, and  $\alpha_0$  signifies the vector of *hyperparameters* needed to specify a member of the parametric family.

A two-level hierarchical model goes one step further,

$$y|\theta, \alpha \sim f(y|\theta, \alpha), \quad \theta|\alpha \sim p_0(\theta|\alpha), \quad \alpha \sim p_1(\alpha) \equiv p_1(\alpha|\beta_0). \quad (3)$$

Here  $p_1$  specifies another parametric family with known hyperparameters  $\beta_0$  and is sometimes called a hyperprior. All of our examples from this section can be written in this form for appropriate definitions of the vectors  $y$ ,  $\theta$ , and  $\alpha$ . Examples 4.12.1 and 4.12.2 both specify  $y = (y_1, \dots, y_m)' \sim f(y|\theta)$ . Moreover  $\theta = (\theta_1, \dots, \theta_m)' \sim p_0(\theta|\alpha)$  where  $\alpha$  is two-dimensional in each case but with very different meanings in the two examples. As in our discussion of Section B.3, from the model statement for observables  $y \sim f(y|\theta)$ , it is implicit that  $y \perp\!\!\!\perp \alpha|\theta$ , so this density can also be written as  $f(y|\theta, \alpha)$ . Example 4.12.4 is similar to Examples 4.12.1 and 4.12.2 but simpler.

Example 4.12.3 has  $y = (y_{11}, \dots, y_{mn_m})'$ , with  $\theta = (\theta_1, \dots, \theta_m)'$  and  $\alpha = (\mu, \alpha_1, \alpha_2)'$ . The model for observable  $y$  is specified by  $y \sim f(y|\theta, \alpha)$ . Again from Section B.3, this means  $y|\theta, \alpha \sim f(y|\theta, \alpha)$  and it is implicit that  $y \perp\!\!\!\perp \mu, \alpha_1|\theta, \alpha_2$  so that  $f(y|\theta, \alpha) = f(y|\theta, \alpha)$ . The model for  $\theta$  has density  $p_0(\theta|\mu, \alpha_1)$ . Implicitly, the unobservables have  $\theta \perp\!\!\!\perp \alpha_2|\mu, \alpha_1$ , so  $p_0(\theta|\mu, \alpha_1) = p_0(\theta|\mu, \alpha_1, \alpha_2) \equiv p_0(\theta|\alpha)$ . Finally,  $\alpha$  is a three-dimensional vector for which we need to specify a prior.

The hierarchical model can be viewed in two different ways. In random effects and latent variable models, the sampling distribution is defined by

$$y|\theta, \alpha \sim f(y|\theta, \alpha), \quad \theta \sim p_0(\theta|\alpha),$$

so that the actual sampling density is

$$f(y|\alpha) = \int f(y|\theta, \alpha)p_0(\theta|\alpha)d\theta \quad (4)$$

and the prior is

$$\alpha \sim p_1(\alpha|\beta_0). \quad (5)$$

The sampling density may be intractable except as a two-stage specification.

Alternatively, the sampling distribution may remain

$$y|\theta, \alpha \sim f(y|\theta, \alpha) \quad (6)$$

with parameters  $\theta$  and  $\alpha$  and the rest of the model is just a specification of the prior

$$p(\theta, \alpha) = p_0(\theta|\alpha)p_1(\alpha|\beta_0). \quad (7)$$

In the special case where

$$y|\theta, \alpha \sim f(y|\theta),$$

the first way of viewing the model (4) and (5) is relatively unchanged, but the alternative view (6) and (7) is that we are using a more general family of priors than  $p_0(\cdot|\cdot)$ . The generalization is specified via

$$\theta \sim p_0(\theta|\alpha), \quad \alpha \sim p_1(\alpha|\beta_0).$$

The actual prior is the mixture

$$p(\theta|\beta_0) = \int p_0(\theta|\alpha)p_1(\alpha|\beta_0)d\alpha$$

but, again, the two-stage specification may be more convenient. For example, the simple mixture of two distributions  $p_0(\theta|\alpha_1)$  and  $p_0(\theta|\alpha_2)$ , say,  $\beta_0 p_0(\theta|\alpha_1) + (1 - \beta_0) p_0(\theta|\alpha_2)$  can be specified by taking  $p_1(\alpha|\beta_0)$  to be a two-point distribution taking the value  $\alpha_1$  with probability  $\beta_0$  and taking the value  $\alpha_2$  with probability  $1 - \beta_0$ . In Chapter 15 we model the  $\theta_i$ s with a Dirichlet Process. This leads to a discrete random mixture distribution for the data, resulting in a non-parametric model for the data.

In either alternative, the structure of the hierarchical model (3) is unchanged. It is only our interpretations of the model that differ. Or perhaps more accurately, it is our modeling intentions that differ.

**EXERCISE 4.27.** The following WinBUGS code corresponds to a hierarchical model for two-stage sampling with Binomial counts and with success probabilities modeled using a Beta distribution. The distribution of success probabilities is called the prevalence distribution.

```
model{
  for(i in 1:k){
    y[i] ~ dbin(theta[i],n[i])
    theta[i] ~ dbeta(alpha1,alpha2)
  }
  mu ~ dbeta(a,b)
  psi ~ dgamma(c,d)
  alpha1 <- mu*psi
  alpha2 <- (1-mu)*psi
  prev ~ dbeta(alpha1,alpha2) # Gets estimate of the
                             # distribution of prevalences
  prob <- step(prev - 0.5) # Pr(prev > 0.5|y)
}
```

Suppose our best guess for  $\mu$  is 0.25 and that we are 95% sure that  $\mu \leq 0.4$ . A Beta( $a, b$ ) distribution with  $a = 8.5$  and  $b = 23.6$  has a mode of 0.25 and has 95% of its area to the left of 0.4.

To obtain a best guess for  $\psi$ , we make a best guess for the entire distribution of prevalences [the Beta( $\alpha_1, \alpha_2$ ) distribution]. We guess the mean to be 0.25 and the 90th percentile to be 0.5. We found that the Beta(1.33, 3.97) satisfies these characteristics. (a) Use these best guesses of  $\alpha_1$  and  $\alpha_2$  to obtain best guesses of  $\mu$  and  $\psi$ . Model your knowledge about  $\psi$  using a Gamma( $c, d$ ) distribution. With best guess for  $\psi$  of  $\psi_0$ , equate  $\psi_0$  to the mode  $(c - 1)/d$  of the Gamma( $c, d$ ) distribution, and

solve for  $c$  in terms of  $d$ . (b) Using the indicated values of  $(a, b, c)$  with  $d = 100, 10, 1, 0.01, 0.001$ , monitor `psi` and `prev` to examine the induced prior on  $\psi$  and the induced “prior prevalence density,” `prev`. Discuss the effect of the choice of  $d$  on these objects. To examine the priors, modify the code by eliminating the specification of the data being binomial.

### 4.13 Some Final Comments on Likelihoods\*

We are so used to specifying models through a sampling distribution  $f(y|\theta)$  and a prior  $p(\theta)$  that it is easy to forget that they are a mere artifice. What is fundamentally important is  $f(y)$ , the distribution of the observable data  $y$ , and the distribution  $f(\tilde{y}|y)$  of future observables  $\tilde{y}$ . We remarked in Section 5 that these are more fundamental quantities than posterior distributions.

Even though the likelihood function has a place of honor in both frequentist and Bayesian statistics, it is a rather artificial construction. If you accept that parameters are artificial constructs, then likelihoods must also be artificial constructs.

In an extension of our discussions on testing, suppose we only really care about a function of the parameters, say  $\gamma \equiv \gamma(\theta)$ . For example, if  $y \sim N(\mu, 1/\tau)$  we may only care about  $\mu$ . Is the appropriate likelihood function

$$L(\theta|y) = f(y|\theta)$$

or is it

$$L(\gamma|y) = f(y|\gamma)?$$

In this latter formula (with only slight abuse of measure theory)

$$f(y|\gamma) = \int_{\gamma(\theta)=\gamma} f(y|\theta)p(\theta|\gamma(\theta)) = \gamma d\theta.$$

In practice, this computation of the conditional distribution would involve a reparameterization of  $\theta$  into  $\tilde{\gamma} \equiv (\gamma', \gamma'_*)'$  with a corresponding transformation of  $p(\theta)$  into  $q(\tilde{\gamma})$  using Proposition B.4. Then, after obtaining the conditional density of  $\gamma_*$  given  $\gamma$ , say  $q(\gamma_*|\gamma)$ , we would compute

$$L(\gamma|y) = f(y|\gamma) = \int f(y|\gamma, \gamma_*) q(\gamma_*|\gamma) d\gamma_*.$$

While this discussion may seem as artificial as the likelihood function it is meant to debunk, the issue has already come up in regard to hierarchical models. Many two-level hierarchical models take the form

$$y|\theta \sim f(y|\theta), \quad \theta|\gamma \sim p_0(\theta|\gamma), \quad \gamma \sim p_1(\gamma|\beta_0).$$

It is by no means clear what the parameters are! Is the sampling distribution

$$y|\theta \sim f(y|\theta)$$

with a prior on  $\theta$  determined by

$$\theta|\gamma \sim p_0(\theta|\gamma), \quad \gamma \sim p_1(\gamma|\beta_0)?$$

Or is the sampling distribution determined by

$$y|\theta \sim f(y|\theta), \quad \theta \sim p_0(\theta|\gamma)$$

with prior

$$\gamma \sim p_1(\gamma|\beta_0)?$$

In the former case, the sampling density is  $f(y|\theta)$  and the parameter vector is  $\theta$ . If the latter case, the sampling density is

$$f(y|\gamma) = \int f(y|\theta)p_0(\theta|\gamma)d\theta$$

and the parameter vector is  $\gamma$ . In the former case, we have a standard sampling distribution with an hierarchical prior. In the latter case, we have a random effects model with a standard prior. Yet the two cases are the same model!

The size of the parameter vectors  $\theta$  and  $\gamma$  can be very different. These differences can have a big effect on model selection criteria like the DIC that penalize a model for the number of parameters it contains. In particular, with DIC it is possible to get different values for different interpretations of the same model. An even bigger dimensionality problem occurs if, as discussed in the previous section,  $y \sim f(y|\theta, \gamma)$  so that

$$f(y|\gamma) = \int f(y|\theta, \gamma)p_0(\theta|\gamma)d\theta.$$

Now the question of the appropriate parameters is no longer between  $\theta$  and  $\gamma$  but between  $\theta$  and  $\gamma$  combined versus  $\gamma$  alone. For the simpler models considered in this book, there is no problem. However, when fitting hierarchical models with random effects (latent variable constructions), the DIC obtained from the marginal likelihood for  $\gamma$  (having integrated out unobservable random effects or latent variables) is different from the DIC for parameters  $\theta$  and  $\gamma$ , obtained by treating random effects/latent variables as parameters with priors (which is standard in WinBUGS). See Spiegelhalter et al. (2002) for more details.

#### 4.14 Identifiability and Noninformative Data

Typically we define our sampling distribution via  $y|\theta$ , the conditional distribution of the observed data given some parameter vector  $\theta$ . This parameter is said to be *identifiable* if the distribution of  $y$  determines the value of the parameter. Technically,  $\theta$  is identifiable if  $y|\theta_1 \sim y|\theta_2$  implies that  $\theta_1 = \theta_2$ .

Conversely,  $\theta$  is not identifiable if there exists  $\theta_1 \neq \theta_2$  with  $y|\theta_1 \sim y|\theta_2$ . Since more than one parameter vector gives exactly the same distribution for the data, we have no hope of determining which parameter is correct from merely observing  $y$ . Even if we somehow managed to observe the entire distribution, we could still not identify the parameter.

The key issue in nonidentifiability is having more parameters than you need to determine the data distribution.

**EXAMPLE 4.14.1.** *One-way Analysis of Variance (ANOVA).* A perfectly well defined model for groups of normal observations indexed by  $i$  is that for  $i = 1, \dots, a$  and  $j = 1, \dots, n_i$ ,

$$y_{ij}|\mu_1, \dots, \mu_a, \tau \stackrel{ind}{\sim} N(\mu_i, 1/\tau).$$

As discussed in Chapters 7 and 9, this “one-way ANOVA” model is often written

$$y_{ij} = \mu_i + \varepsilon_{ij}; \quad \varepsilon_{ij} \stackrel{iid}{\sim} N(0, 1/\tau).$$

Not infrequently, the model is written in an overparameterized version

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}; \quad \varepsilon_{ij} \stackrel{iid}{\sim} N(0, 1/\tau), \tag{1}$$

where  $\mu$  is referred to as the grand mean and  $\alpha_i$  is referred to as the effect for group  $i$ . In model (1), the data vector  $y$  consists of all the  $y_{ij}$  observations and  $\theta = (\mu, \alpha_1, \dots, \alpha_a, \tau)'$ .  $\theta$  is not identifiable. For example  $\theta_1 = (0, \mu_1, \dots, \mu_a, \tau)'$  and  $\theta_2 = (2, \mu_1 - 2, \dots, \mu_a - 2, \tau)'$  determine the same distribution for  $y$ .

Identifiability problems often arise because we specify a distribution for potential observables  $Z$  but then actually observe only some function of  $Z$ . Suppose  $Z$  depends on some natural parameter

vector  $\theta$  that is identifiable. However, rather than observing  $Z$ , we observe a function of  $Z$ , say,  $y = h(Z)$ . It is this observable random vector  $y$  for which  $\theta$  is often nonidentifiable, i.e., for some  $\theta_1 \neq \theta_2$  we have  $y|\theta_1 \sim y|\theta_2$ .

EXAMPLE 4.14.2. *Mixture Models.* A perfectly well-defined model has

$$X_k \stackrel{\text{ind}}{\sim} N(\mu_k, 1), k = 1, 2 \quad \perp\!\!\!\perp \quad W \sim \text{Bern}(p).$$

Let  $Z = [X_1, X_2, W]'$ . The parameter vector is  $\theta' = [\mu_1, \mu_2, p]$ . As discussed further in Chapter 11 and Section 15.1, in a mixture model what we actually see is a function of  $Z$ ,

$$y = WX_1 + (1 - W)X_2.$$

The two distinct parameter vectors  $\theta'_1 = [0, 2, 0.3]$  and  $\theta'_2 = [2, 0, 0.7]$  determine the same distribution for  $y$ . Both have a 30% chance of coming from a  $N(0, 1)$  and a 70% chance of coming from a  $N(2, 1)$ .

Often there is a function of the parameter vector  $\theta$ , say  $g(\theta)$ , that is an identifiable parameter for  $y$ . In other words, the distribution  $y|g(\theta)$  is uniquely determined by  $g(\theta)$ . In the one-way ANOVA problem,  $g_1(\theta) = (0, \alpha_1, \dots, \alpha_a, \tau)', g_2(\theta) = (\mu, 0, \alpha_2, \dots, \alpha_a, \tau)', g_3(\theta) = (\mu, \alpha_1, \dots, \alpha_{a-1}, -\alpha_1 - \dots - \alpha_{a-1}, \tau)',$  and  $g_4(\theta) = (\mu + \alpha_1, \dots, \mu + \alpha_a, \tau)'$  are all identifiable parameterizations. The first three of them simply restrict the permissible values for the parameters. The third one is a restriction that  $\alpha_1 + \dots + \alpha_a = 0$ . The fourth function actually changes the dimension of the parameterization from  $a+2$  to  $a+1$ . The mixture of two normal models can be made identifiable either by incorporating the restriction that  $\mu_1 < \mu_2$  or by incorporating the restriction that  $p > 0.5$ .

EXAMPLE 4.14.3. *Drug Testing.* We extend Example 2.2.1 using slightly different notation than we later use in Chapter 14. If you use drugs ( $D$ ), you can either test positive for them ( $T^+$ ) or test negative ( $T^-$ ). Non-users are denoted  $\bar{D}$ . In a sample of  $n$  people, the numbers in each categorization are denoted

Data	$D$	$\bar{D}$
$T^+$	$z_{11}$	$z_{12}$
$T^-$	$z_{21}$	$z_{22}$

An obvious parameterization for these data consists of the probabilities for each category, say  $p_{ij}$ . The probabilities must be nonnegative and add to one. An alternative parameterization is based on the prevalence of drug use in the population,  $\pi = \Pr(D) = p_{11} + p_{21}$ , the sensitivity of the drug test,  $\eta = \Pr(T^+|D) = p_{11}/\pi$ , and the specificity of the drug test,  $\xi = \Pr(T^-|\bar{D}) = p_{22}/(1 - \pi)$ . In particular,

Probabilities	$D$	$\bar{D}$
$T^+$	$p_{11} = \eta\pi$	$p_{12} = (1 - \xi)(1 - \pi)$
$T^-$	$p_{21} = (1 - \eta)\pi$	$p_{22} = \xi(1 - \pi)$

For the full data  $Z = (z_{11}, z_{12}, z_{21}, z_{22})'$ , the restricted  $p_{ij}$ s and  $\theta = (\pi, \eta, \xi)'$  are identifiable parameters. However, it is often the case that we do not have a way of identifying who uses drugs and who does not, so the only data we see are the data on who tests positive. In this case,

$$y = z_{11} + z_{12}.$$

It is easy to see that

$$y \sim \text{Bin}(n, \pi\eta + (1 - \pi)(1 - \xi)),$$

so the *apparent prevalence*, defined as  $g(\theta) = \pi\eta + (1 - \pi)(1 - \xi)$ , is an identifiable parameter for  $y$ .

Whether  $g(\theta)$  arises naturally, like the apparent prevalence, or we have to construct it, as in the other examples,  $y$  gives information on  $g(\theta)$  but not on all of  $\theta$ . In particular, the distribution of  $\theta$  given  $g(\theta)$  is the same in the posterior as in the prior, i.e.,

$$\theta|g(\theta), y \sim \theta|g(\theta), \quad (2)$$

see also the end of Section B.3.

For drug testing, in Chapter 14 we define a prior on  $\theta = (\pi, \eta, \xi)'$  and we would like the posterior on  $\theta$ , i.e., the posterior on the sensitivity, specificity, and prevalence. These are the parameters we really care about. In that case, relation (2) is disturbing, i.e., the distribution of the sensitivity, specificity, and prevalence given the apparent prevalence does not depend on the data. On the other hand, for ANOVA models or the mixture models of Section 15.1, we would be very happy to define the prior directly on  $g(\theta)$  and restrict our posterior inferences to  $g(\theta)$ .

In practice, we often define a prior on  $\theta$  and induce a prior on  $g(\theta)$ . For example, in the our two normals mixture model, a general yet convenient prior is

$$\mu_1, \mu_2 \stackrel{iid}{\sim} N(\tilde{m}, 1/\tilde{\tau}) \quad \perp \!\!\! \perp p \sim \text{Beta}(\tilde{\alpha}, \tilde{\alpha}).$$

This prior can be problematic but the problems are solved by forcing the means to satisfy  $\mu_1 < \mu_2$ . This amounts to giving prior probability one to the means being ordered, so we need worry no more about restricting the means. (More on this in Subsection 15.1.1.)

*One problem with Bayesian analysis is that it is easy to overlook identifiability issues.* Given a prior  $p(\theta)$ , you can incorporate the data through  $f(y|\theta)$  and get a posterior without ever realizing that

$$f(y|\theta) \equiv f(y|g(\theta))$$

and that (2) is going to keep you from learning about all aspects of  $\theta$ .

Finally, we consider data that provide no information on some aspects of the parameters. Suppose the data have distribution  $y|\theta$  and suppose  $\theta$  can be reparameterized into  $\xi = (\xi'_1, \xi'_2)'$ . If we can write  $L(\xi|y) = q(\xi_1|y)$ , then these specific data  $y$  contain no information about the parameters  $\xi_2$ . Any prior on  $\theta$  induces a prior on  $\xi$ , and with a likelihood of this form  $\xi_2|\xi_1, y \sim \xi_2|\xi_1$ . In particular, if  $\xi_1$  and  $\xi_2$  are independent in the prior, the posterior distribution of  $\xi_2$  is independent of  $\xi_1$  and exactly the same as its prior.

This is precisely what happens when  $y|\theta$  is not identifiable but  $y|g(\theta)$  is. In that case, we can reparameterize  $\theta$  as  $\xi$  where  $\xi_1 = g(\theta)$ . In the nonidentifiable case, no matter what data you see, there will be no information on the parameters  $\xi_2$  or equivalently,  $\theta|g(\theta), y \sim \theta|g(\theta)$  regardless of the value of  $y$ . However, in other situations, notably when we deal with the proportional hazards model in Section 13.2, the specific value of  $y$  determines which parameters  $\xi_1$  have information and which parameters  $\xi_2$  have no information.

---

## Chapter 5

---

# Comparing Populations

---

At long last we now get serious about analyzing data. Chapter 1 introduced typical results from Bayesian data analysis. Chapter 2 covered fundamental scientific and mathematical ideas behind the Bayesian approach. Chapter 3 showed how to avoid the nasty calculus associated with Bayesian computations. Chapter 4 discussed a wide array of statistical concepts useful in Bayesian statistics. In particular, Chapters 2 and 4 discussed one sample data for Bernoullis, binomials, and normals with known precision. Example 3.1.3 also discussed two independent binomial samples. But all of those discussions focused on illustrating concepts of Bayesian statistics. Now we focus on analyzing data from one or two populations. Methods are presented for commonly used parametric models including the binomial, negative binomial, normal with unknown precision, and Poisson. In particular, Section 1 discusses inferences for proportions, rates and effect measures under binomial, negative binomial and case-control sampling. Section 2 discusses one- and two-sample normal populations. Section 3 discusses one- and two-sample Poisson data. All sections contain details of how to specify prior distributions. Material on analyzing one and two samples for time to event data (survival analysis/reliability analysis) appears in Chapter 12.

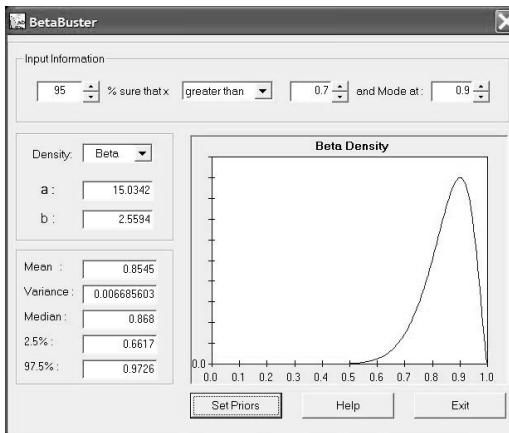
Bayesian analysis requires prior distributions. Priors can involve substantive scientific input or they can be chosen as convenient reference priors. For discrete data, our reference priors for standard one- and two-sample problems often include Jeffreys' priors, c.f. Section 4.7, or suitable approximations to them. Our substantive priors are often taken as conjugate priors. Scientific input is obtained by eliciting information about objects that scientists can easily think about. This input is then applied to the parameters of the model.

Having specified a likelihood and a prior, the posterior is found either analytically (if convenient) or by generating Monte Carlo samples from the posterior. Throughout, we look for interesting population characteristics that are readily available, that is, available by simply writing down a logical expression for the parameter(s) of interest.

### 5.1 Inference for Proportions

Does smoking increase the probability of lung cancer? Do Fords break down more than Toyotas? Would a new treatment better solve your medical problem than the conventional therapy? Which of two major airlines, Red-eye Express and Fly BiNite, has a higher proportion of on-time arrivals? These questions involve comparing probabilities for two groups. Relevant data involve sampling individuals from each group (people, cars, patients, flights) and recording binary outcomes (cancer, breakdown, cure, on-time). More complex problems with binomial data are examined in Chapter 8. In Chapter 1 we considered a simple example with only one group (population).

**EXAMPLE 5.1.1.** In Section 1.1 we considered data from the St. Paul Brass and Aluminum Foundry. Out of 2,430 push rod eyes poured at their foundry only 2,211 actually shipped. Our interest is in the probability of pouring a defective part. The Vice President of Operations thinks that a plausible range for the proportion of scrap is 5% to 15%.

Figure 5.1: *BetaBuster GUI*.

We assume that the number of defective parts  $y$  has a binomial distribution

$$y|\theta \sim \text{Bin}(2430, \theta).$$

We need a prior on  $\theta$ . We interpreted the Vice President as specifying  $\Pr(\theta \leq 0.05) = 0.025$  and  $\Pr(\theta \geq 0.15) = 0.025$ . Assuming that  $\theta$  is modeled with a Beta prior, these two probability statements can be shown to determine the Beta distribution

$$\theta \sim \text{Beta}(12.05, 116.06),$$

see Exercise 5.1.

With  $y = 219 = 2430 - 2211$ , as discussed in Example 2.3.1 the exact posterior distribution is Beta(219 + 12.05, 2430 - 219 + 116.06) with median 0.0902 and 95% probability interval (0.0795, 0.1017). Computer simulations gave the estimated median of the posterior distribution as 0.09 with 95% PI for  $\theta$  of (0.08, 0.10).

Finding the Beta distribution from two probability statements as in Example 5.1.1 is not trivial. A procedure for determining the Beta parameters  $a$  and  $b$  from the mode and one upper or lower percentile of the distribution has been automated in a program named *BetaBuster* that is freely available to download at <http://www.epi.ucdavis.edu/diagnostictests/>. *BetaBuster* computes the parameters, reports summary values of the distribution, and provides a graph of the density. Figure 5.1 illustrates the *BetaBuster* GUI. Exercise 5.1 illustrates that *BetaBuster* can be used to determine the Beta prior for Example 5.1.1.

**EXERCISE 5.1.** BetaBuster requires that you give it the mode and a percentile of the distribution and it returns information, including the parameters, on the Beta distribution corresponding to those values. Enter one percentile from Example 5.1.1, say, that 0.05 is the 2.5th percentile of the Beta. Start with an initial guess that the mode is 0.1 = (0.05 + 0.15)/2 and check the 97.5 percentile of the Beta as reported by BetaBuster. By trial and error, adjust the mode so that the 97.5 percentile is 0.15. Report your values of the Beta parameters  $a$  and  $b$ .

**EXERCISE 5.2.** Use the following code to reproduce the results of Example 5.1.1.

```
model{
y ~ dbin(theta,n)
theta ~ dbeta(12.05,116.06)
```

```

}
list(n=2430,y=219)
list(theta=0.5)

```

### 5.1.1 Prior Distributions

We now discuss prior distributions for binomial data. The Beta family provides a flexible and convenient class of distributions for modeling uncertainty about probabilities. The density is given in Table 2.1. Different choices for the parameters  $a$  and  $b$  lead to a variety of density shapes including U-shaped, J-shaped, L-shaped, unimodal symmetric, unimodal skewed right and left, and the  $U(0, 1)$  density. No parametric family is broad enough to capture all the possibilities for a prior distribution on  $\theta$ . Finding a prior that agrees with the information of the expert is far more important than the convenience of using a Beta distribution. For example, if the expert's prior is bimodal (rare in our experience), we must abandon Beta distributions because they are only bimodal when  $a, b < 1$  (with modes of 0 and 1). We could use a mixture of Betas, however, which is discussed below. In our experience a single Beta has worked well for most problems.

We begin by discussing reference priors, only because we have relatively little to say about them. Most of this subsection is devoted to methods of eliciting substantive information for use in a Beta prior. The subsection closes with some discussion of mixtures of Beta priors and truncated Beta priors.

#### 5.1.1.1 Reference Priors

For a binomial random variable  $y|\theta \sim \text{Bin}(n, \theta)$  there is little agreement on how to choose a reference prior from among the three standard candidates: (1) the improper  $\text{Beta}(0, 0)$  distribution discussed in Example 4.6.3, (2) Jeffreys' prior  $\text{Beta}(0.5, 0.5)$  discussed in Example 4.7.2, and (3) the  $U(0, 1) = \text{Beta}(1, 1)$  prior. Geisser (1984) discusses the relative merits of these priors. The first two put most of their probability on values very near 0 and 1, the improper prior overwhelmingly so. (We told you “noninformative priors” might be stupid!) Fortunately, these two priors tend to have little effect on the posterior when sample sizes are moderate and when the data don't concentrate near 0 or 1. Tuyl et al. (2008a) point out unfortunate consequences when the data are all successes or all failures. All three priors are data augmentation priors in the sense that they can be viewed as adding prior successes and failures to the data. The improper prior adds no prior observations, Jeffreys' prior adds one prior observation with half a success and half a failure, and the uniform adds two observations with one success and one failure. See also Tuyl, Gerlach, and Mengersen (2008b).

#### 5.1.1.2 Informative Beta Priors

Most often we obtain a prior on the probability  $\theta$  from an expert by eliciting two pieces of information:

- I. Their best guess for the probability.
- II. The biggest value the probability could *reasonably* be, say, a value with only a 5% chance that the probability would exceed it.

More generally, our elicitation involves the expert guessing the value of  $\theta$  and providing a percentile for the distribution of  $\theta$ . We typically treat the best guess as the mode of the prior distribution. Beta-Buster was designed to provide the parameters of the Beta distribution given two such inputs. Given the circumstances, what we ask for may be the smallest number the probability could reasonably be and we might change the “chance” (the percentage) associated with reasonableness. If the prior guess of  $\theta$  is less than 0.5, we typically ask for an upper value that the expert is 95% or 99% sure  $\theta$  falls below. If the prior guess is greater than 0.5, we typically ascertain a lower value that the expert thinks  $\theta$  exceeds with high probability. These values must be elicited independently of the current data being analyzed. Priors are often elicited after the current data have been collected which makes

elicitation ignoring the current data more difficult. Rather than asking an expert, these inputs could be based on published values or historical data.

BetaBuster uses a fairly simple idea. Suppose that our prior mode is  $\theta_0$ . A Beta( $a, b$ ) distribution has mode

$$\theta_0 = \frac{a - 1}{a + b - 2}$$

for  $a, b > 1$ , so solving we get

$$a(b) = \frac{1 + \theta_0(b - 2)}{1 - \theta_0}. \quad (1)$$

Now we can find  $b$  by using information on a percentile, say a percentile  $c$  such that

$$\alpha = \Pr[0 < \theta < c] = \int_0^c \frac{\Gamma(a(b) + b)}{\Gamma(a(b))\Gamma(b)} \theta^{a(b)-1} (1 - \theta)^{b-1} d\theta.$$

A simple computerized search procedure allows us to find  $a$  and  $b$ . Suppose  $\theta_0 = 0.2$ ,  $\alpha = 0.95$ , and  $c = 0.45$ . We create a vector of possible  $b$  values,  $\tilde{b} = (1, \dots, 100)'$  and corresponding  $a(b)$  values. Then we find the 0.95 quantiles (95th percentiles) for all of these Beta( $a(b), b$ ) distributions, say,  $qbeta(a(b), b)$ , and check the quantiles against the specified prior quantile  $c = 0.45$ . Suppose we find that  $0.45 \in (qbeta(a(10), 10), qbeta(a(11), 11))$ . We create a new vector  $\tilde{b}1 = 10 + (\tilde{b}/100)$ , and repeat the process using  $\tilde{b}1$  in place of  $\tilde{b}$ .

The following R commands execute this procedure. The first code tries values of  $b \in \{1, 2, \dots, 100\}$ ; the second looks at  $b1 \in \{10.01, 10.02, \dots, 11\}$ .

```
# Code 1                                # Code 2
t0 <- 0.2                                b <- 1:100
b <- 1:100                                 t0 <- 0.2
a <- (1 + t0*(b-2))/(1-t0)                b1 <- 10+(b/100)
qbeta(0.95, a, b)                          a <- (1 + t0*(b1-2))/(1-t0)
                                            qbeta(0.95, a, b1)
```

The result is a Beta(3.31, 10.24) distribution. BetaBuster gives the distribution as Beta(3.3094, 10.2374).

### 5.1.1.3 Rare Events

We now consider priors for very rare events. When dealing with probabilities that are very small (or large), we actually have a great deal of prior information, and we ignore or lessen that information at our own peril.

Suppose we are 95% sure that  $\theta < 0.10$ , and we believe that values of  $\theta$  closer to zero are more plausible than those that are not. Then a reasonable choice of prior is a Beta distribution with a mode of 0 and 95% of the area under the density to the left of 0.10. If we pick  $a = 1$  and  $b > 1$ , it is easy to see from the form of the Beta density that the mode will be 0. (If  $b = 1$  and  $a > 1$ , the mode is 1.)

With  $a = 1$  and  $b > 1$  simple calculus lets us solve for the  $b$  that gives  $\Pr[\theta < 0.10] = 0.95$ . By the definition of the Gamma function,

$$\Gamma(b + 1) = b\Gamma(b); \quad \Gamma(1) = 1.$$

With  $c = 0.10$  and the Beta density from Table 2.1

$$\begin{aligned} 0.95 &= \int_0^c p(\theta)d\theta \\ &= \int_0^c \frac{\Gamma(1+b)}{\Gamma(1)\Gamma(b)} \theta^{1-1} (1-\theta)^{b-1} d\theta \\ &= \int_0^c b(1-\theta)^{b-1} d\theta \end{aligned}$$

$$\begin{aligned} &= -(1 - \theta)^b \Big|_0^c \\ &= 1 - (1 - c)^b \end{aligned}$$

Solving for  $b$  gives

$$b = \frac{\log(1 - 0.95)}{\log(1 - c)},$$

which is 28.43 for  $c = 0.10$ . A Beta(1, 28.43) agrees with our prior beliefs. If we take an arbitrary percentile  $\alpha$ ,  $\alpha = \int_0^c p(\theta)d\theta$  and  $b = \log(1 - \alpha)/\log(1 - c)$ .

**EXAMPLE 5.1.2.** Tuyl et al. (2008a) discuss potential dangers of using priors having  $a < 1$  with data that are all zeros (or  $b < 1$  with all ones). We broaden the discussion to inherent difficulties with specifying priors in situations where it is known *a priori* that the parameter value is at or near the boundary of the parameter space. Specifically, Tuyl et al. 2008a consider data on “bad reactions” to a *new* radiological contrast agent. Let  $\theta_N$  be the probability of a bad reaction using the *new* agent. The *standard* agent causes bad reactions about 15 times in 10,000 which, if the new agent is similar to the standard, suggests a “best guess” for the new agent probability  $\theta_N$  of 0.0015. Following Tuyl et al. 2008a consider two priors for this probability, both with mean 0.0015; a Beta(1, 666) and a Beta(0.05, 33.33). Both priors have mode 0 which suggests an implicit belief that values are increasingly plausible for decreasing values of  $\theta_N$ . (Technically, the Beta(0.05, 33.33) does not have a mode by our definition, but the density goes to infinity as  $\theta$  approaches 0.) We think of the first prior as 1 prior “bad reaction” in 667 prior trials and the second as 0.05 prior “bad reactions” in 33.38 prior trials. The number of prior trials indicates the amount of information in the prior. The prior probabilities that  $\theta_N \leq 0.0015$  are 0.63 and 0.88 for the two priors. The 95th percentiles are 0.0045 and 0.0081, respectively. These priors have consequences and the scientist needs to be aware of them before analyzing data.

We consider two sets of hypothetical data based on using the *new* agent, each with 100 trials, the first has no “bad reactions” and the second has one. We also consider two additional priors that have mode and median equal to 0.0015, respectively. With a prior belief that the new agent is similar to the standard, all four priors for  $\theta_N$  are “centered” on information known for the standard agent. We look for evidence that the proportion of reactions under the new agent is less than 0.0015. Results for the various probabilities that  $\theta_N \leq 0.0015$  follow:

Centering	Pr[ $\theta < 0.0015$ ]			
	Mean	Mean	Mode	Median
( $a, b$ )	(1, 666)	(0.05, 33.33)	(1.6, 407.4)	(1.05, 497)
Prior	0.632	0.882	0.222	0.5
Post. (0/100)	0.683	0.939	0.289	0.568
Post. (1/100)	0.319	0.162	0.074	0.213

As expected, results from the mean centered priors indicate that the posterior probability of  $\theta_N \leq 0.0015$  is changed more by the two data sets when using the low weight prior.

Imagine how much information we must have on the standard agent to be able to say with confidence that reactions occur about 15 times in 10,000. To change to the new agent, we should require data strong enough to make an impact on those very strong prior beliefs. As a design issue, one would probably want to continue sampling until *at least* one reaction occurs. With the standard agent, we expect to see  $667 = 1/0.0015$  trials before getting a reaction. Intuitively we would want a lot more than 100 observations (with no reactions) before we claimed that the new agent was better. Yet with the low weight prior, one is 94% sure that the new agent is better after only 100 “good” trials. It turns out that it takes 1,330 trials without a reaction to be 95% sure that the new agent is better using the higher weight Beta(1, 666) prior, whereas only 145 good trials are needed with the Beta(0.05, 33.33) prior. The low weight prior largely assumes the conclusion that is hoped for, thus it takes inappropriately little confirmatory data to reach that conclusion.

Alternatives to mean centered priors are the mode and median centered priors. In each case we used 0.01 as an effective upper bound for  $\theta_N$  but with the mode centered prior we took 0.01 as the 95th percentile and with the median centered prior we took it to be the 99th percentile. Inferences are all over the place depending on the prior specification, despite the fact that all priors are “centered” at 0.0015. The median and mode centered priors both have non-zero modes and, under each, the plausibility for values of  $\theta_N$  tend to zero as we move to the left. The mode based prior has only a 0.22 prior probability of being less than 0.0015, is not moved very much by seeing no bad reactions, and drops to 0.074 when 1 bad reaction is seen out of 100. The moral is that we *want* a prior with a lot of weight so that it takes a lot of data to change it. No small amount of data, like 100 observations, will be convincing. In problems where the posterior is not overly concentrated near the boundary of the parameter space (zero in this case), centering at the mean, median, and mode will be much more similar than they are here.

A more appropriate prior might be to use one that is not conjugate. If one believes that  $\theta_N$  must be very small, a  $U(0, u)$ , might be appropriate for some  $u$ . If  $u = 0.01$ , for example, this reflects a prior certainty that the rate/proportion of bad reactions could never be above 100/10,000, and that there is indifference to all possible values below this number. With this prior and with 0 bad reactions out of 100, the posterior probability that  $\theta_N \leq 0.0015$  is 0.22. With a  $U(0, 1)$  prior, the corresponding probability is 0.14; the median of the  $U(0, 1)$  based posterior for  $\theta_N$  is 0.0069 versus 0.0037 in the  $U(0, 0.01)$  analysis. The estimated probabilities of the new agent being better are of similar orders of magnitude, the estimates of  $\theta_N$  are not; the  $U(0, 1)$  prior effectively adds one “success” and one “failure” to data with 0 “failures” and 100 “successes” and thus moves the estimate up. The  $U(0, 1)$  prior is actually dis-informative.

Finally, all of the priors that focus on small values of  $\theta_N$  are presuming that the new agent will work similarly to the standard agent. It would probably be safer to argue that either the new agent will work like the standard agent or that we have little idea of how it will work. As such, a more appropriate prior for this problem might well be a mixture of a prior that focuses on small values of  $\theta_N$  and a reference prior such as the  $U(0, 1)$ .

A clear message from this analysis is that in the absence of huge amounts of data, the choice of prior matters a lot when prior information (and ultimately the data) dictate that the parameter is near the boundary of the parameter space. In particular, a Beta prior with  $a \leq 1$  or  $b \leq 1$  must be followed with very careful consideration of its implications.

**EXERCISE 5.3.** Using calculus, find the mode and 5th percentile of a Beta(10,1) distribution.

**EXERCISE 5.4.** Using calculus, find  $a$  and  $b$  such that a Beta( $a, b$ ) distribution has a mode of 1 and a 5th percentile of 0.2.

**EXERCISE 5.5.** Derive formula (1), including the formula for  $\theta_0$ .

**EXERCISE 5.6.** Use BetaBuster to find the Beta( $a, b$ ) priors for mode 0.75 and 5th percentile 0.60, and for mode 0.01 and 99th percentile 0.02. What is the Beta prior when the mode is 1 and the first percentile is 0.80?

**EXERCISE 5.7.** The distributions  $\theta \sim \text{Beta}(1.6, 1)$  and  $\theta \sim \text{Beta}(1, 0.577)$  both have a mode of 1. Find  $\Pr[\theta < 0.5]$  analytically for each. Does BetaBuster give the appropriate parameters for the Beta distributions?

#### 5.1.1.4 Non-Beta Priors

What do I do if my prior is bimodal? The Beta distribution won’t work. For example, suppose  $\theta$  is the probability that a person is taller than six feet and you know that the data will come from a population that includes either all men or all women, but you do not know which. Your prior

should be a combination (mixture) of your prior on  $\theta$  for women and your prior on  $\theta$  for men. The combination should weight these individual distributions by your prior probability that the sampled population will be all women (or men). Alternatively, a bimodal prior might arise as a combined prior from two experts having very different theories or opinions on the same issue. In Example 5.1.2 we suggested a possible mixture prior because our prior information might not be relevant to the problem at hand.

A simple solution to this problem is to use a mixture of Beta distributions. For example, if we let  $\text{Beta}(\cdot|a,b)$  denote the density of a  $\text{Beta}(a,b)$  distribution, we could use a prior

$$p(\theta) = w\text{Beta}(\theta|a_1,b_1) + (1-w)\text{Beta}(\theta|a_2,b_2),$$

where  $a_1, a_2, b_1, b_2$  are known and  $w \in [0, 1]$ . Depending on prior information,  $w$  can be known, e.g., the prior probability that the data on being over six feet are from a population of women, or we can consider this as just a more general family of distributions and let  $w$  be an unknown parameter. In the later case, we would place a prior on  $w$ , possibly another Beta, giving us a hierarchical prior as discussed in Section 4.12. For example, this might be appropriate if  $w$  reflects a third expert's opinion about the veracity of the theories held by two other experts. Fortunately, bimodal prior distributions are relatively uncommon.

Another prior that arises occasionally is restricted to a subset of  $[0,1]$ . For example, if a casino gives one dollar for winning a one dollar bet, we can be 100% sure that  $\theta < 0.5$ . In Example 5.1.2 we discussed a prior that assumed the rate of bad reactions was below 0.01. Generally, if  $0 \leq s < \theta < t \leq 1$ , we can consider a density  $p(\theta) \propto \text{Beta}(\theta|a,b)I_{(s,t)}(\theta)$ . This distribution is a truncated Beta. Choosing  $a = b = 1$  results in a  $U(s,t)$  prior.

**EXERCISE 5.8.** Write WinBUGS code to simulate several choices of Beta distributions for different choices of  $(a,b)$ . Consider a variety with  $a$  or  $b < 1$ ,  $= 1$ , and  $> 1$ . Then using the code below, simulate a 50-50 mixture of  $\text{Beta}(10,20)$  and  $\text{Beta}(20,10)$  random variables. Modify the code and try some of your own choices for Beta distributions and mixing weight  $w$ . Note that when you are in WinBUGS, there are no data to input, so you simply skip the step of inputting data and proceed in the usual way after compiling.

```
model{
  gamma[1] ~ dbeta(10,20)
  gamma[2] ~ dbeta(20,10)
  theta <- w*gamma[1] + (1-w)*gamma[2]
  w ~ dbern(0.5)
}
```

**EXERCISE 5.9.** Write WinBUGS code to simulate from  $\text{Beta}(\theta|a,b)I_{(s,t)}(\theta)$  densities using several different choices of  $(a,b)$  and  $(s,t)$ . This is done by writing

```
theta ~ dbeta(a,b)I(s,t)
```

in WinBUGS. Although truncated Beta distributions are not common, an appropriate Markov chain will be obtained. If  $s = 0$  or if  $t = 1$ , use  $I(,t)$  or  $I(s,)$ , respectively. In WinBUGS, one can also write uniform distributions as `dunif(s,t)`. Note that you must specify an initial value for `theta` that is in  $(s,t)$ .

### 5.1.2 Effect Measures

With binary data the main subject of interest is the probability of “success”  $\theta$ . In biosciences  $\theta$  is often referred to as a *risk*. For example, the proportion of smokers who develop lung cancer denotes the “risk” of smoking. The word “risk” is used to connote both good and bad outcomes just as we call  $\theta$  the probability of success even though  $\theta$  may be the probability of death or failure.

A related measure of risk is the odds. The odds  $O$  are the ratio of the success probability to the failure probability,

$$O = \theta / (1 - \theta).$$

We can retrieve the probability from the odds because

$$\theta = O / (1 + O).$$

With  $0 < \theta < 1$ , the odds is a positive number, i.e.,  $0 < O < \infty$ . The odds gets larger as  $\theta$  gets larger. Things that cannot happen have both zero probability and odds of zero. Odds of one correspond to  $\theta = 0.5$ . The odds is infinite when the probability is one, i.e., the event is a sure thing. Why would anyone abandon a perfectly intuitive measure like probability to discuss odds? Especially a measure that is a single number but pretends to be plural? You'd have to ask your bookie. Sometimes we look at the log of the odds. Log-odds take values from negative infinity to positive infinity with  $\theta = 0.5$  corresponding to log-odds of zero. This property of log-odds becomes useful in Chapters 7 and 8.

Life gets more complicated and more interesting when we have two populations to compare. We can compare the probabilities of lung cancer for smokers, say  $\theta_1$ , and non-smokers,  $\theta_2$ . One comparison looks at their ratio, known as the *relative risk (RR)*,

$$RR \equiv \theta_1 / \theta_2.$$

A relative risk (*risk ratio*) can also be the ratio of probabilities for, say, learning to juggle in three hours using two different teaching methods. It is a very general concept. An *RR* of 3 means that the event of interest is three times more likely in the numerator group than in the other.

For two populations we may also look at the *odds ratio* defined as

$$OR = \frac{\theta_1 / (1 - \theta_1)}{\theta_2 / (1 - \theta_2)}.$$

Of course we may also take the logarithm of this. Why would anyone abandon a perfectly intuitive measure like relative risks to discuss odds ratios? You'd have to ask your mathematical statistician. Oh, that's us. Well, mathematically, odds ratios have some very nice properties. But we doubt that you want us to expound on the relationships between log-odds ratios and interaction contrasts, cf. Christensen (1997, Subsection 2.5.1). We will also see in Chapters 7 and 8 that log-odds ratios appear naturally in many of our models.

Suppose  $OR = 2$ . That occurs when  $\theta_1 = 1/2$  and  $\theta_2 = 1/3$ . It also occurs when  $\theta_1 = 2/7$  and  $\theta_2 = 1/6$  and in many other situations. It *almost* occurs when  $\theta_1 = 0.002$  and  $\theta_2 = 0.001$ . The risk ratio and the odds ratio are very similar numbers when both of the risks are small, i.e., both probabilities are near zero. ORs are commonly used in epidemiology to study rare diseases in which case they approximate the RR. ORs are necessarily the parameters of interest when data are collected from a case-control study. Case-control studies are discussed in Subsection 5.1.4.

A third effect measure for two populations is the *risk difference*,

$$RD = \theta_1 - \theta_2,$$

also called the *attributable difference*. For smoking this is the difference between the proportion of smokers who get lung cancer and the proportion of non-smokers who get lung cancer.

Although odds and odds ratios are staples of medical and social research, we believe that *RR* and *RD* are easier to interpret, so preference should go to making statistical inferences about them. Point estimates and probability intervals for *RD*, *RR*, or *OR* are readily available using a sample from the posterior. They can be used jointly to assess both *statistical* and *practical* differences between two probabilities. For example, if a 99% PI for *RR* is (9.9, 10.1), we are virtually certain that one risk is essentially ten times the other. This would be both practically and statistically important. On the

other hand, if the 99% PI was (1.01, 1.02), we would be virtually certain that one risk was larger than the other, but would anyone care?

**EXERCISE 5.10.** Find three sets of  $(\theta_1, \theta_2)$  values not given earlier that correspond to  $OR = 2$ . Give the corresponding  $RR$  and  $RD$  values. Argue that  $OR = RR$  when the  $\theta$ s are close to zero.

### 5.1.3 Independent Binomials

We now consider data analyzed as two independent binomials.

**EXAMPLE 5.1.3.** *Comparing Binomial Proportions.* Carey et al. (1998) studied the “cost of reproduction” on longevity of Mediterranean fruit flies (medflies). We use data from the 534 medflies that lived at least 34 days. These are cross-classified by two factors: A) whether the fly produced at least 1,000 eggs in its first 30 days of life and B) whether the fly lived at least 44 days, see Table 5.1.

Table 5.1: *Medfly data.*

Early Egg Production	Longevity	
	Long-lived	Short-lived
High	54	80
Low	224	176

These are clearly multinomial data but since our interest is in using the reproductive factor to explain the response factor longevity, it makes sense to condition the observations on their observed reproductive factor. This allows us to treat the number of long-lived flies as two independent binomial distributions,

$$y_1|\theta_1 \sim \text{Bin}(n_1, \theta_1) \quad \perp \quad y_2|\theta_2 \sim \text{Bin}(n_2, \theta_2).$$

Here,  $\theta_1$  and  $\theta_2$  denote the probability of long lifetimes for high early-life egg producers and low early-life egg producers, respectively. The actual data are  $n_1 = 134$ ,  $n_2 = 400$ ,  $y_1 = 54$ , and  $y_2 = 224$  where the  $n_i$ s are viewed as fixed known numbers and the  $y_i$ s are viewed as observations on random variables. The sample proportion of long-lived medflies is 0.40 for high early producers and 0.56 for low early producers.

As in Example 3.1.3, if we assume independent  $\text{Beta}(a_i, b_i)$  priors for  $\theta_i$ ,  $i = 1, 2$ , we obtain the posterior distribution

$$\theta_1|y_1 \sim \text{Beta}(y_1 + a_1, n_1 - y_1 + b_1) \quad \perp \quad \theta_2|y_2 \sim \text{Beta}(y_2 + a_2, n_2 - y_2 + b_2).$$

Note that the posterior of  $\theta_1$  does not depend on  $y_2$  etc.

Due to faulty financial oversight, our budget for eliciting expert opinions on this medfly example was accidentally spent on a Club Med membership. In the absence of available prior information we place independent  $U(0, 1) = \text{Beta}(1, 1)$  reference priors on the probabilities. (See Exercise 5.12 for an analysis with informative priors.) The posteriors are

$$\theta_1|y_1 \sim \text{Beta}(55, 81) \quad \perp \quad \theta_2|y_2 \sim \text{Beta}(225, 177).$$

The posterior means are  $E(\theta_1|y_1) = 55/(55+81) = 0.40$  and  $E(\theta_2|y_2) = 225/(225+177) = 0.56$ . From Table 5.2, the posterior medians, 95% PIs are 0.40, (0.32, 0.49) for the high early producers and 0.56, (0.51, 0.61) for the low early producers. The interval for the high producers is entirely below the interval for the low producers suggesting that longevity is increased for flies with lower early egg production. Similarly, Table 5.2 shows strong evidence that the relative risk and odds ratio are both greater than one and that the risk difference is greater than zero. Medflies with a lower amount of early egg production are estimated to be about 1.38 times more likely to be long-lived compared to high early producers and we are 95% sure that the actual value is in the interval

Table 5.2: Posterior summaries for medfly data.

Parameter	mean	sd	2.5%	median	97.5%
$\theta_1$	0.405	0.042	0.324	0.404	0.487
$\theta_2$	0.560	0.025	0.511	0.560	0.607
$RR$	1.400	0.162	1.124	1.384	1.750
$OR$	1.915	0.393	1.266	1.873	2.802
$RD$	0.155	0.049	0.059	0.155	0.250
$I_{(0,\infty)}(RD)$	0.999	0.031	1.0	1.0	1.0

(1.12, 1.75). The mean in the last row of Table 5.2, 0.9991, gives the posterior probability that  $RD$  is at least zero. Clearly, medflies that lead a dissipated youth tend to die an early death.

EXERCISE 5.11. Reproduce the results of Example 5.1.3 using the following WinBUGS code.

```
model{
  y[1] ~ dbin(theta[1], n[1])
  y[2] ~ dbin(theta[2], n[2])
  theta[1] ~ dbeta(1, 1)
  theta[2] ~ dbeta(1, 1)
  odds[1] <- theta[1]/(1-theta[1])
  odds[2] <- theta[2]/(1-theta[2])
  RD <- theta[2]-theta[1]
  RR <- theta[2]/theta[1]
  OR <- odds[2]/odds[1]
  test <- step(RD)
}
list(n=c(134, 400), y=c(54, 224))
list(theta=c(0.5, 0.5))
```

The step function was discussed near the end of Section 3.2.

EXERCISE 5.12. Suppose an expert's best guess for  $\theta_1$  is 0.3 and they are 95% sure that  $\theta_1$  is less than 0.6. Their best guess for  $\theta_2$  is 0.5 and they are 95% sure that it is less than 0.9. Re-analyze the medfly data using this prior information and compare results to the reference analysis given above.

In Example 4.3.3 we compared binomial to negative binomial sampling. In binomial sampling we have a fixed number of Bernoulli trials and count the random number of successes. In negative binomial sampling we have a sequence of Bernoulli trials that stops when we hit a fixed number of successes, thus the number of trials is random. The parameter, the Bernoulli probability of success,  $\theta$  is the same in either case, so prior information is obtained the same way in either case. As discussed earlier, with the same prior and proportional likelihoods we get the same Bayesian inferences from binomial and negative binomial sampling.

EXERCISE 5.13. A study was conducted to compare the proportions of televisions A and B that are defective immediately after manufacturing. The TVs were sampled independently until exactly 20 defective As and 20 defective Bs were found. TV A required  $y_1 = 59$  trials to reach 20 defectives and B required  $y_2 = 179$ . Let  $\theta_1$  and  $\theta_2$  denote the population proportions of defective As and Bs, respectively.

- (a) Suppose prior information is extracted from a consumer magazine that has previously reported on the chance of buying a defective A or B. The prior estimate of the probability of a defective A is 0.1 and 0.05 for B. Also suppose we are 99% sure that less than 20% of each of the types of TVs are defective. Construct independent Beta priors that embody this information.

- (b) Give the exact posterior distribution of  $(\theta_1, \theta_2)$  and find the posterior means, modes, standard deviations, and (using R) 95% probability intervals. Use the prior from part (a).
- (c) Write a WinBUGS program to analyze the data. Obtain posterior means, medians, standard deviations, and 95% intervals for  $\theta_1$ ,  $\theta_2$ ,  $RD$ , and  $RR$ . Interpret your results.
- (d) Test whether  $\theta_2 > \theta_1$ . Provide a conclusion.

#### 5.1.4 Case-Control Sampling

The strategy used in Example 5.1.3 is a common one. We sampled a number of individuals (medflies) and cross-classified them by an explanatory factor and a response factor. We then conditioned on the explanatory factor to arrive at data from two independent binomials. Alternatively, if the explanatory factor defines convenient populations, we can take independent samples from each level of the explanatory factor. In our example, high early egg producers and low early egg producers do not define medfly subpopulations that are easy to sample.

Both of these sampling strategies break down when studying rare events. Suppose we wanted to study the relationship between childhood vaccinations and autism. Autism is still relatively rare, so we would need a very large sample of children if we were to include, say, 50 autistic children. This would be true regardless of whether we sample children and cross-classify them, or whether we could conveniently sample the subpopulations of vaccinated children and unvaccinated children (a highly unlikely prospect).

Rather than either of these sampling techniques, we might conveniently take one sample of autistic children and one sample of nonautistic children, that is, a sample of cases and a sample of controls. We could then classify children in each group by whether they had received childhood vaccinations. If we have a population of autistic children available, it becomes much easier to ensure that the study includes a reasonable number of them.

**EXAMPLE 5.1.4.** In the late 1970s, it was observed that, in a sample of  $n_1 = 7$  children with Reye's Syndrome (RS), all 7 of them were taking aspirin at the time they became sick. A second sample of size  $n_2 = 16$  children known to be free of Reye's Syndrome was also taken, and it was determined that 8 of them were taking aspirin when sampled. For more details on these data, see Gastwirth (1988, Vol. 2).

The MLE of the probability of a child being on aspirin in the RS group is 1 and the corresponding estimate in the non-RS group is 0.5. Based on these estimates  $\widehat{OR} = \infty$ . Just for fun, let's see what happens if we had seen 6 aspirin takers with RS instead of the actual 7. The MLE of the *OR* would be  $\widehat{OR} = [(6/7)/(1/7)]/[(1/2)/1/2] = 6$ , which epidemiologists would normally think of as quite large. But with the small sample sizes, is the result statistically important? More importantly, why would we care? Why would we care that children with RS are more likely to take aspirin? Isn't what we care about whether children who take aspirin are more likely to get RS?

The problem with case-control sampling for vaccinations and autism is that we want to study  $\tilde{\theta}_i$ s, the probabilities that children are autistic given their vaccination status, but what the data allow us to study are  $\theta_j$ s, the probabilities of being vaccinated given their autism status. Although we cannot study the  $\tilde{\theta}_i$ s directly, we will exploit, and later show, that the odds ratio based on the  $\tilde{\theta}_i$ s is the same as the odds ratio based on the  $\theta_j$ s, a value that we *can* estimate.

To put it another way, assume a disease of interest  $D$  with  $D = 1$  indicating presence and  $D = 2$  indicating absence of the disease and an exposure variable  $E$  with  $E = 1$  indicating exposure and  $E = 2$  indicating no exposure. For example,  $E = 1$  corresponds to "vaccinated" and  $E = 2$  corresponds to "non-vaccinated" while  $D = 1$  corresponds to "autism" and  $D = 2$  corresponds to "no autism." We are interested in determining how  $E$  affects the probability of  $D$ , namely  $\Pr(D|E)$ . We want to compare  $\Pr(D = 1|E = 1) \equiv \tilde{\theta}_1$  and  $\Pr(D = 1|E = 2) \equiv \tilde{\theta}_2$ . Ideally, we would like to know, say, the risk ratio

$$RR = \Pr(D = 1|E = 1)/\Pr(D = 1|E = 2) = \tilde{\theta}_1/\tilde{\theta}_2,$$

i.e., the proportion of autism among vaccinated children relative to the proportion of autism among non-vaccinated children. Alternatively, we might be interested in the risk difference or even the least intuitive of the effect measures, the odds ratio.

Unfortunately, our study design does not allow us to estimate the risk ratio, or the risk difference, or the  $\theta_i$ s. However, the odds ratio of being autistic when vaccinated relative to being autistic when not vaccinated turns out to be estimable.

With our study design, we can estimate

$$\theta_1 \equiv \Pr(E = 1|D = 1) \quad \text{and} \quad \theta_2 \equiv \Pr(E = 1|D = 2).$$

Moreover, the odds ratio of being autistic when vaccinated relative to being autistic when not vaccinated turns out to be a function of the  $\theta_j$ s. In particular, we will show that the odds ratio we want turns out to equal the odds ratio of being vaccinated when autistic relative to being vaccinated when not autistic, that is,

$$OR = \frac{\theta_1/(1-\theta_1)}{\theta_2/(1-\theta_2)} = \frac{\tilde{\theta}_1/(1-\tilde{\theta}_1)}{\tilde{\theta}_2/(1-\tilde{\theta}_2)}.$$

If we are willing to look at odds ratios as our effect measure, we can examine the relative effect of vaccination on the odds of autism even from a case-control study. This is not exactly what we hoped to achieve; we wanted to estimate the risk ratio. But when looking at small probabilities, like the probability of autism, the odds ratio is a good approximation to the risk ratio. For example if  $OR = 10$ , the odds of autism in the vaccinated group would be 10 times that for the non-vaccinated group. If  $\Pr(D = 1|E = 1) = 0.01$  and  $\Pr(D = 1|E = 2) = 0.001$  we get  $RR = 10 \doteq OR$ . Unfortunately, when the probabilities are not small, say if  $\Pr(D = 1|E = 1) = 10/19 = 0.53$  and  $\Pr(D = 1|E = 2) = 0.1$ ,  $OR = 10$  but  $RR = 5.3$ , a big difference.

The general model for our simple case-control study consists of two independent binomials  $y_1$  and  $y_2$  that give the numbers of  $E = 1$  individuals from the cases and controls, respectively. We assume that independently

$$y_i|\theta_i \sim \text{Bin}(n_i, \theta_i).$$

The Bayesian approach to analysis takes on different forms depending on how the prior information is specified. If the experts have prior information on  $\theta_1 = \Pr(E = 1|D = 1)$  and  $\theta_2 = \Pr(E = 1|D = 2)$ , then simple independent Beta priors on these may suffice and the analysis is just like the others illustrated in this section. The only caveat is that interest focuses almost exclusively on the odds ratio.

**EXERCISE 5.14.** (a) Analyze the Reye's Syndrome data using  $U[0,1]$  priors on  $\theta_1$  and  $\theta_2$ . (b) Discuss issues of causation and correlation for the RS data. For example, people with high blood pressure are more likely to take beta blocker drugs like atenolol than people without high blood pressure. Can we conclude that using beta blockers causes high blood pressure? Does using BetaBuster cause high blood pressure?

Another situation that can arise is where there is prior information on  $OR$ , perhaps based on a previous case-control study that is similar to the one at hand. We can place a prior on the parameters by placing, say, a normal prior on  $\delta \equiv \log(OR)$  and an independent Beta prior on  $\theta_2$ . These induce a prior on  $(\theta_1, \theta_2)$ . To apply this, we need to solve for  $\theta_1$  in terms of  $\delta$  and  $\theta_2$ . Some algebra gives

$$\theta_1 = \frac{e^\delta \theta_2}{1 - \theta_2(1 - e^\delta)}.$$

To elicit a prior on  $\delta$  we think about  $OR$ . If our best guess is that  $OR = 3$ , then we take the mean of the normal distribution for  $\delta$  to be  $\log(3) = 1.1$ . Moreover, if we are, say, 90% sure that the  $OR$  is at least 0.8, then we are also 90% sure that  $\log(OR)$  is at least  $\log(0.8) = -0.22$ . We need to find a normal distribution with a mean of 1.1 and a 10th percentile of -0.22. We know (or can look

up) that the 10th percentile of a normal is  $-1.28$  standard deviations below the mean, so we set  $1.1 - 1.28\sigma = -0.22$  and solve for  $\sigma = 1.03$ . Our prior on  $\delta$  is  $N(1.1, (1.03)^2)$ .

**EXERCISE 5.15.** An expert has no belief that there is or is not an effect of aspirin on RS, but that if there is one, it won't be "huge" in either direction. They place a  $N(0, 2)$  prior on  $\delta$ , so their best guess for the *OR* is  $e^0 = 1$ . The prior also indicates that they are 95% sure that the *OR* is in the interval  $e^{(-1.96*1.414, 1.96*1.414)} = (0.063, 16.0)$ . The interval is "centered" on 1 and allows for a broad range of possibilities in both directions. Place the Jeffreys' Beta(0.5, 0.5) prior on  $\theta_2$ .

Analyze the RS data by adding to the following code.

```
model{
  for(i in 1:2){ y[i] ~ dbin(theta[i], n[i]) }
  theta[2] ~ dbeta(a, b)
  delta ~ dnorm(mu, prec)
  theta[1] <- exp(delta)*theta[2]/(1-theta[2]*(1-exp(delta)))
  OR <- theta[1]/(1-theta[1])/((theta[2]/(1-theta[2])))
}
```

Examine the sensitivity of the results to the choice of prior. Try a prior that reflects much more skepticism about whether there is any effect and a prior that suggests that any effect will be a positive one. Also consider a  $U[\log(0.02), \log(50)]$  prior for  $\delta$  and a Beta(1, 1) prior for  $\theta_2$  to see what impact that has on the results. Be sure to calculate the posterior probability that  $OR > 1$ , and possibly some other posterior probabilities, like the probability that  $OR > 2$ .

A third scenario involves eliciting prior information on  $\tilde{\theta}_1 = \Pr(D = 1|E = 1)$ ,  $\tilde{\theta}_2 = \Pr(D = 1|E = 2)$ , and  $\gamma \equiv \Pr(E = 1)$ . With prior information on these three parameters, we can combine it with the information from the case-control data to obtain posterior inferences for all parameters of interest including the risk ratio of the  $\tilde{\theta}_j$ s. To do this we need to define the relationship between the  $\tilde{\theta}_j$ s and the  $\theta_j$ s. Using Bayes' Theorem,

$$\theta_1 = \frac{\tilde{\theta}_1 \gamma}{\tilde{\theta}_1 \gamma + \tilde{\theta}_2 (1 - \gamma)}; \quad \theta_2 = \frac{(1 - \tilde{\theta}_1) \gamma}{(1 - \tilde{\theta}_1) \gamma + (1 - \tilde{\theta}_2) (1 - \gamma)}. \quad (2)$$

The prior on  $(\tilde{\theta}_1, \tilde{\theta}_2, \gamma)$  induces a prior on  $(\theta_1, \theta_2)$ .

Of course there is a catch! There is no information in our case-control data about the probability that someone would have autism. More generally case-control data contain no information about

$$\Pr[D = 1] = \tilde{\theta}_1 \gamma + \tilde{\theta}_2 (1 - \gamma).$$

Thus there exists a function of the parameters used in the prior whose distribution will not change in the posterior. What we can learn about from the case-control data are  $\theta_1$  and  $\theta_2$  and anything that is a function of them, like the *OR*. If we try to make inferences on anything else, like the *RR* for the  $\tilde{\theta}_j$ s, there is no guarantee that more data will get us closer to the "true" answer because aspects of the posterior will depend only on the prior and not the data. For example,

$$\tilde{\theta}_1 = \frac{\theta_1 \Pr[D = 1]}{\theta_1 \Pr[D = 1] + \theta_2 (1 - \Pr[D = 1])}.$$

The data inform us about the  $\theta_j$ s but not about  $\Pr[D = 1]$  so we cannot be sure that more data will give us the true  $\tilde{\theta}_1$ . In particular, the conditional prior distribution of  $\tilde{\theta}_1, \tilde{\theta}_2, \gamma$  given  $\theta_1, \theta_2$  will be identical to the conditional posterior distribution of  $\tilde{\theta}_1, \tilde{\theta}_2, \gamma$  given  $\theta_1, \theta_2, y_1, y_2$ . This last approach to analyzing case-control data is closely related to Chapter 14 on diagnostic testing.

**EXERCISE 5.16.** In Exercise 5.15, modify the problem so that you have:

- (a) Beta priors on  $(\tilde{\theta}_1, \tilde{\theta}_2, \gamma)$  where  $\tilde{\theta}_1$  and  $\tilde{\theta}_2$  have modes of 0.7 and 0.3, and the 5th and 95th percentiles, respectively, both equal 0.5. Let  $\gamma \sim \text{Beta}(2, 2)$ . Analyze the data and compare inferences for this and the analyses of the two previous exercises.
- (b) Repeat part (a) only now fix the percentiles for the  $\tilde{\theta}_j$ s at 0.6 and 0.4, respectively.

**EXERCISE 5.17.** (a) For the prior in Exercise 5.15, simulate the induced prior on the  $\tilde{\theta}_j$ s or give a reason why not. (b) For the priors in Exercise 5.16 (a) and (b), simulate the induced prior on the  $\theta_j$ s.

**EXERCISE 5.18.**

- (a) Re-visit Exercise 5.14 where the Reye's Syndrome (RS) data were analyzed. Suppose follow-up case-control data were taken from the same two populations of RS cases and non-RS controls. Using your posterior from the previous analysis, construct a prior for the current analysis, and then analyze the current data wherein the number of RS cases sampled was 37 of which 35 were taking aspirin, and out of 75 controls, 32 were taking aspirin.
- (b) Perform a sensitivity analysis on the prior. Would you be willing to conclude that the probability of RS is greater for those on aspirin than for those not on aspirin? Justify.

Finally, we show that the odds ratio of being vaccinated when autistic relative to being vaccinated when not autistic is the same as the odds ratio of being autistic when vaccinated relative to being autistic when not vaccinated. Recall that the  $\tilde{\theta}_j$ s are the probabilities that children are autistic given their vaccination status, but case-control data allow us to study only the  $\theta_j$ s, the probabilities of being vaccinated given their autism status. In a slight modification of (2), the probability of being vaccinated when autistic is

$$\theta_1 = \Pr[E = 1|D = 1] = \frac{\Pr[D = 1|E = 1]\Pr[E = 1]}{\Pr[D = 1]} = \frac{\tilde{\theta}_1\Pr[E = 1]}{\Pr[D = 1]}.$$

The odds of being vaccinated when autistic are

$$\frac{\theta_1}{1 - \theta_1} = \frac{\Pr[E = 1|D = 1]}{\Pr[E = 2|D = 1]} = \frac{\Pr[D = 1|E = 1]\Pr[E = 1]}{\Pr[D = 1|E = 2]\Pr[E = 2]} = \frac{\tilde{\theta}_1\Pr[E = 1]}{\tilde{\theta}_2\Pr[E = 2]}.$$

Similarly, the probability of being vaccinated when not autistic is

$$\theta_2 = \Pr[E = 1|D = 2] = \frac{\Pr[D = 2|E = 1]\Pr[E = 1]}{\Pr[D = 2]} = \frac{(1 - \tilde{\theta}_1)\Pr[E = 1]}{\Pr[D = 2]}$$

and the odds of being vaccinated when not autistic are

$$\frac{\theta_2}{1 - \theta_2} = \frac{\Pr[E = 1|D = 2]}{\Pr[E = 2|D = 2]} = \frac{\Pr[D = 2|E = 1]\Pr[E = 1]}{\Pr[D = 2|E = 2]\Pr[E = 2]} = \frac{(1 - \tilde{\theta}_1)\Pr[E = 1]}{(1 - \tilde{\theta}_2)\Pr[E = 2]}.$$

Finally, the odds ratio is

$$OR = \frac{\theta_1/(1 - \theta_1)}{\theta_2/(1 - \theta_2)} = \frac{\tilde{\theta}_1\Pr[E = 1]}{\tilde{\theta}_2\Pr[E = 2]} \Big/ \frac{(1 - \tilde{\theta}_1)\Pr[E = 1]}{(1 - \tilde{\theta}_2)\Pr[E = 2]} = \frac{\tilde{\theta}_1/(1 - \tilde{\theta}_1)}{\tilde{\theta}_2/(1 - \tilde{\theta}_2)}.$$

Thus, the odds ratio of being vaccinated when autistic relative to vaccinated when not autistic is the same as the odds ratio of being autistic when vaccinated relative to being autistic when not vaccinated.

## 5.2 Inference for Normal Populations

Normal data are ubiquitous in the scientific world because they often arise from measurements. We've all been exposed to the "bell-shaped curve." Standardized test scores displayed in a histogram look like a bell. Data on how long people can hold their breath under water look like a diving bell.

**EXAMPLE 5.2.1.** In Section 1.2 we discussed a corrosion-resistant brass alloy used in plumbing fixtures. The addition of zinc to the alloy produces a jump in metal strength. Interest lies in the percentage of zinc so that it can be adjusted to be within specifications. Twelve alloy samples were tested with  $y_i$  denoting the zinc percentage of sample  $i$ ,  $i = 1, \dots, 12$ . Assume

$$y_1, \dots, y_{12} | \mu, \sigma^2 \stackrel{iid}{\sim} N(\mu, \sigma^2).$$

The  $n = 12$  sample zinc percentages are 4.20, 4.36, 4.11, 3.96, 5.63, 4.50, 5.64, 4.38, 4.45, 3.67, 5.26, and 4.66. The sample mean is  $\bar{y} = 4.568$ , the sample standard deviation is  $s = 0.631$ , so the standard error of  $\bar{y}$  is  $SE(\bar{y}) = 0.631/\sqrt{12} = 0.182$ . The minimum observation is  $y_{(1)} = 3.67$  and the maximum is  $y_{(12)} = 5.64$ . The first quartile is  $q_1 = 4.133$ , the third is  $q_3 = 5.110$ , and the median is  $\tilde{y} = 4.415$ . The data seemed reasonably normal when inspecting a normal plot.

Let  $\tau \equiv 1/\sigma^2$  and suppose we have observations

$$y_1, \dots, y_n | \mu, \sigma^2 \stackrel{iid}{\sim} N(\mu, 1/\tau).$$

We need to place a prior on  $\theta \equiv (\mu, \tau)'$ . Traditionally, this was done using either a reference prior or a conjugate prior. We find the conjugate prior to have limited practical utility for data analysis because we find it difficult to elicit expert information for the parameters. We focus primarily on priors with independent information on  $\mu$  and  $\tau$ .

### 5.2.1 Reference Priors

The most commonly used reference prior puts "independent" flat priors on both  $\mu$  and  $\log(\tau)$ , i.e.,

$$p(\mu, \tau) = \frac{1}{\tau},$$

see Section 4.6. We refer to this prior as the *standard improper reference (SIR) prior*. Its posterior results mimic the usual frequentist results for normal data. *We think of this as a justification for using the usual frequentist results.* With the sample mean and sample variance defined as

$$\bar{y} = \sum_{i=1}^n y_i/n; \quad s^2 = \sum_{i=1}^n (y_i - \bar{y})^2/(n-1),$$

at the end of this subsection we establish that

$$\mu | \tau, y \sim N(\bar{y}, 1/n\tau) \tag{1}$$

and

$$\tau | y \sim \text{Gamma}[(n-1)/2, (n-1)s^2/2]. \tag{2}$$

Frequentist confidence intervals are identical to Bayesian posterior intervals under this prior. In particular, using Exercise 5.19 below, the marginal distribution of  $\tau$  is determined by

$$(n-1)s^2 \tau | y = \frac{(n-1)s^2}{\sigma^2} | y \sim \chi_{n-1}^2.$$

If we let  $\ell = \chi_{n-1}^2(\alpha/2)$  and  $u = \chi_{n-1}^2(1 - \alpha/2)$ , then we have

$$1 - \alpha = \Pr[\ell \leq (n-1)s^2\tau \leq u | y] = \Pr\left[\frac{(n-1)s^2}{u} \leq \sigma^2 \leq \frac{(n-1)s^2}{\ell} \middle| y\right].$$

If  $\alpha = 0.05$ , we are 95% sure that  $(n-1)s^2/u \leq \sigma^2 \leq (n-1)s^2/\ell$ . This is the Bayesian analogue of the standard frequentist confidence interval for  $\sigma^2$  based on the distribution of  $s^2$  given the variance  $\sigma^2$ . However, in the Bayesian analysis,  $s^2$  is fixed and  $\sigma^2$  (or  $\tau$ ) is treated as random.

**EXERCISE 5.19.** Let  $v \sim \text{Gamma}(a, bc)$ . Show that  $bv \sim \text{Gamma}(a, c)$ . Thus, since a  $\text{Gamma}(k/2, 1/2)$  variate is also a  $\chi^2(k)$  variate, we must have  $2bcv \sim \chi^2(2a)$  provided  $2a$  is an integer.

Exercise 5.19 provides a way to define  $\chi_a^2$  for noninteger values of  $a > 0$ , i.e.,

$$\chi^2(a) \equiv \text{Gamma}\left(\frac{a}{2}, \frac{1}{2}\right).$$

This result will be needed in Subsection 5.2.2.

As shown later, the marginal posterior distribution of  $\mu$  is determined by

$$\frac{\mu - \bar{y}_\cdot}{\sqrt{s^2/n}} \middle| y \sim t(n-1). \quad (3)$$

This looks similar to the standard frequentist distributional result for random  $\bar{y}_\cdot$  and  $s^2$  with given mean  $\mu$ ; however, in the Bayesian analysis,  $\bar{y}_\cdot$  and  $s^2$  are fixed and  $\mu$  is random. Given this result, it is easy to see that a 95% probability interval for  $\mu$  has

$$0.95 = \Pr\left[\bar{y}_\cdot - t(0.975, n-1)\sqrt{s^2/n} \leq \mu \leq \bar{y}_\cdot + t(0.975, n-1)\sqrt{s^2/n} \middle| y\right].$$

This has exactly the same form as the frequentist confidence interval, except that it is perfectly proper now to discuss the probability of  $\mu$  being in the interval — something it is nearly impossible to stop introductory students from doing. More generally, a  $1 - \alpha$  PI has endpoints:  $\bar{y}_\cdot \pm t(1 - \alpha/2, n-1)\sqrt{s^2/n}$ .

The correspondence with frequentist results even carries to the predictive distribution because a new observation  $y_{n+1}$  has

$$\frac{y_{n+1} - \bar{y}_\cdot}{s\sqrt{1+(1/n)}} \middle| y \sim t(n-1).$$

Here the only random variable is the future observation  $y_{n+1}$ , see Exercise 5.21.

**EXAMPLE 5.2.1 CONTINUED.** For the zinc percentages, the posterior distribution of  $\mu$  devolves from

$$\frac{\mu - 4.568}{0.182} \middle| y \sim t(11)$$

and for  $\sigma^2$  from

$$\frac{11(0.631)^2}{\sigma^2} \middle| y \sim \chi^2(11).$$

**EXERCISE 5.20.** For the zinc data, give the 95% PIs for  $\mu$ ,  $\sigma$ , and  $y_{13}$ .

More recently priors such as

$$\mu \sim N(0, b) \quad \perp \!\!\! \perp \quad \tau \sim \text{Gamma}(c, c)$$

have been used as proper reference priors. The key point is that the prior precision for  $\mu$ ,  $1/b$ , and both hyperparameters  $c$  in the Gamma distribution are near zero. Common choices are  $b = 10^6$  and  $c = 0.001$ . Such priors can be viewed as approximations to the  $p(\mu, \tau) = 1/\tau$  SIR prior. The posterior analysis is a special case of Subsection 5.2.3.

#### *Posterior Derivation\**

We now derive the posterior using the SIR prior  $p(\mu, \tau) = 1/\tau$ . Let  $N(\cdot|a, b)$  be a normal density and  $\text{Gamma}(\cdot|c, d)$  be a gamma density. Using Exercise 4.9 and the analysis of Example 4.4.2

$$\begin{aligned} L(\mu, \tau|y) &\propto \tau^{n/2} \exp\left[-\frac{\tau}{2}(n-1)s^2 - \frac{\tau}{2}n(\bar{y} - \mu)^2\right] \\ &= \tau^{1/2} \exp\left[-\frac{n\tau}{2}(\mu - \bar{y})^2\right] \tau^{(n-1)/2} \exp\left[-\frac{(n-1)s^2}{2}\tau\right]. \end{aligned} \quad (4)$$

To get the posterior, multiply by the prior  $p(\mu, \tau) = 1/\tau$  giving

$$p(\mu, \tau|y) \propto \tau^{1/2} \exp\left[-\frac{n\tau}{2}(\mu - \bar{y})^2\right] \tau^{\frac{n-1}{2}-1} \exp\left[-\frac{(n-1)s^2}{2}\tau\right].$$

To obtain the conditional density of  $\mu$  given  $\tau$  and  $y$ , drop the last two multiplicative terms that do not include  $\mu$ ,

$$p(\mu|\tau, y) \propto \tau^{1/2} \exp\left[-\frac{n\tau}{2}(\mu - \bar{y})^2\right] \propto N(\mu|\bar{y}, 1/n\tau).$$

While this is only a function of  $\mu$ , we have written it so that the constant of proportionality does not depend on  $\tau$ . The role of  $\tau$  in the normal density is important in the next step.

To get the marginal posterior distribution of  $\tau$ , in the joint posterior integrate out terms involving  $\mu$ . The normal density integrates to 1, so

$$\begin{aligned} p(\tau|y) &= \int p(\mu, \tau|y)d\mu \\ &\propto \tau^{\frac{n-1}{2}-1} \exp\left[-\frac{(n-1)s^2}{2}\tau\right] \int N(\mu|\bar{y}, 1/n\tau)d\mu \\ &\propto \text{Gamma}\left(\tau \left| \frac{n-1}{2}, \frac{(n-1)s^2}{2}\right.\right). \end{aligned}$$

It follows from Exercise 5.19 that  $(n-1)s^2\tau|y \sim \chi_{n-1}^2$ .

Finally, to get the marginal posterior of  $\mu$ , rewrite

$$\begin{aligned} p(\mu, \tau|y) &\propto \left[\tau^{1/2} e^{-\frac{n\tau}{2}(\mu - \bar{y})^2}\right] \left[\tau^{[(n-1)/2]-1} \exp\left\{-\frac{\tau}{2}(n-1)s^2\right\}\right] \\ &= \tau^{n/2-1} \exp\left\{\frac{-\tau[n(\mu - \bar{y})^2 + (n-1)s^2]}{2}\right\} \end{aligned}$$

which, as a function of  $\tau$ , is proportional to the density of a  $\text{Gamma}(n/2, [n(\mu - \bar{y})^2 + (n-1)s^2]/2)$  distribution. The marginal for  $\mu|y$  is

$$p(\mu|y) = \int_0^\infty p(\mu, \tau|y)d\tau \propto \int_0^\infty \tau^{n/2-1} \exp\left\{\frac{-\tau[n(\mu - \bar{y})^2 + (n-1)s^2]}{2}\right\} d\tau,$$

where we can evaluate the integral because we know that Gamma densities integrate to 1. Evaluating the integral gives

$$\begin{aligned} p(\mu|y) &\propto \frac{\Gamma(n/2)}{[(n-1)s^2 + n(\mu - \bar{y})^2]^{n/2}} \\ &\propto [1 + (\mu - \bar{y})^2/(n-1)(s^2/n)]^{-[(n-1)+1]/2}. \end{aligned}$$

Referring to Table 2.1, we see that as a function of  $\mu$  this is the kernel of a univariate Student density, namely  $t(n - 1, \bar{y}_., \sqrt{s^2/n})$ . Standardizing the Student random variable, that is, subtracting the location and dividing by the dispersion, we obtain a standard  $t(n - 1)$  distribution as in (3).

**EXERCISE 5.21.** Show that the predictive density of  $y_{n+1}$  given  $y$  is  $t(n - 1, \bar{y}_., \sqrt{s^2(1 + 1/n)})$ . Standardizing the  $t$ , derive a 95% prediction interval formula. Compare it with the corresponding prediction interval formula from a frequentist Statistics book.

### 5.2.2 Conjugate Priors

Conjugate priors for normal data with unknown precision are determined by

$$\tau \sim \text{Gamma}\left(\frac{a}{2}, \frac{b}{2}\right); \quad \mu | \tau \sim N(\mu_0, 1/\omega_0\tau).$$

Here  $a$ ,  $b$ ,  $\mu_0$ , and  $\omega_0$  are known numbers chosen to characterize the prior information. Our problem with using this prior in practical data analysis is the difficulty of specifying a distribution for  $\mu$  that is conditional on  $\tau$ .

Before presenting the posterior, define a Bayesian point estimate for  $\mu$  as

$$\hat{\mu}_B = \frac{n}{n + \omega_0} \bar{y}_. + \frac{\omega_0}{n + \omega_0} \mu_0,$$

which is a weighted average of the prior mean and the sample mean. Also define the Bayesian sum of squares error as

$$BSSE = \sum_{i=1}^n (y_i - \bar{y}_.)^2 + \frac{n\omega_0}{n + \omega_0} (\mu_0 - \bar{y}_.)^2.$$

Finally, using the  $\chi^2$  distribution with non-integer degrees of freedom defined after Exercise 5.19, that is,  $\chi_a^2 \equiv \text{Gamma}(a/2, 1/2)$ , correspondingly define a  $t(a, \theta, \sigma)$  distribution simply by replacing the  $n$  in the  $t(n, \theta, \sigma)$  density of Table 2.1 with  $a$ .

The joint posterior,  $p(\mu, \tau | y) = p(\mu | \tau, y)p(\tau | y)$ , is obtained through

$$\mu | \tau, y \sim N(\hat{\mu}_B, 1/\tau(n + \omega_0))$$

and

$$\tau | y \sim \text{Gamma}\left(\frac{n+a}{2}, \frac{BSSE+b}{2}\right).$$

Similar to the previous subsection, the marginal posterior distribution of  $\sigma^2$  (or  $\tau$ ) is determined by

$$(BSSE + b)\tau | y = \frac{BSSE + b}{\sigma^2} | y \sim \chi^2(n + a).$$

It follows from the fact that the mean of a  $\chi^2(n + a)$  is  $n + a$  that  $E(\tau | y) = 1/\hat{\sigma}_B^2$  where  $\hat{\sigma}_B^2 \equiv (BSSE + b)/(n + a)$ . The marginal posterior distribution of  $\mu$  is determined by

$$\frac{\mu - \hat{\mu}_B}{\sqrt{\hat{\sigma}_B^2/(n + \omega_0)}} | y \sim t(n + a).$$

This leads to the 95% probability interval

$$0.95 = \Pr\left[\hat{\mu}_B - t(0.975, n + a)\hat{\sigma}_B/\sqrt{n + \omega_0} \leq \mu \leq \hat{\mu}_B + t(0.975, n + a)\hat{\sigma}_B/\sqrt{n + \omega_0} | y\right].$$

The predictive distribution is determined by

$$\frac{y_{n+1} - \hat{\mu}_B}{\hat{\sigma}_B \sqrt{1 + \frac{1}{n+\omega_0}}} \Big| y \sim t(n+a).$$

This leads to the 95% prediction interval

$$0.95 = \Pr \left[ \hat{\mu}_B - t(.975, n+a) \hat{\sigma}_B \sqrt{1 + \frac{1}{n+\omega_0}} \leq y_{n+1} \leq \hat{\mu}_B + t(0.975, n+a) \hat{\sigma}_B \sqrt{1 + \frac{1}{n+\omega_0}} \Big| y \right].$$

Note that the results on SIR priors *nearly* agree with these when  $\omega_0 = a = b = 0$ . The *only* difference is that when  $\omega_0 = 0$ , as opposed to  $\omega_0 > 0$  but small, the BSSE is the sum of  $n$  terms instead of  $n+1$  terms which causes the first hyperparameter of the gamma distribution to be  $n-1$  rather than  $n+a=n+0$ .

**EXERCISE 5.22.** Derive the posterior distribution for the conjugate prior by using the *complete the square formula*

$$r(\mu - v)^2 + s(\mu - w)^2 = (r+s)(\mu - \hat{\mu})^2 + \frac{rs}{r+s}(v-w)^2, \quad \hat{\mu} = \frac{r}{r+s}v + \frac{s}{r+s}w,$$

and adapting the arguments illustrated in the previous subsection for the SIR prior. (a) Derive the conditional density  $p(\mu | \tau, y)$ . Show that it is the normal density given in this subsection. (b) Using the result in (a), obtain the marginal densities  $p(\mu | y)$  and  $p(\tau | y)$ . (c) Derive the predictive density for a future observation based on this model.

### 5.2.3 Independence Priors

Here we assume that information about  $\mu$  can be elicited independently of information on  $\tau$  or  $\sigma$ , so

$$p(\mu, \tau) = p(\mu)p(\tau).$$

This makes elicitation relatively easy. Although the primary goal is to get a prior that reasonably captures the expert's information, our experience is that independence priors work well.

In Subsection 5.2.2, we focused on Gamma priors for  $\tau$  because they are conjugate. Once we abandon conjugate priors, there is no reason to restrict attention to Gamma priors on  $\tau$ . We can use other priors with positive support. We can also use priors defined on  $\sigma$  or  $\sigma^2$ .

Using independent priors constitutes our first real case in which the posterior cannot be obtained using calculus. Sampling from the posterior is mandatory. Details of posterior sampling for normal data with independent priors are given in Example 6.3.1.

**EXAMPLE 5.2.1 CONTINUED.** We need to identify a prior distribution that gives information about the unknown parameters  $\mu$  and  $\tau = 1/\sigma^2$ . Our expert estimated with 95% certainty that the mean zinc percentage  $\mu$  before pouring should be between 4.5% and 5% and would be centered at 4.75%. We interpret this information as  $E(\mu) = 4.75$  and  $\Pr(4.5 < \mu < 5) = 0.95$  in the prior distribution. Further assuming that the prior on  $\mu$  is normal, we find

$$\mu \sim N(4.75, 0.0163).$$

(Determining the prior variance of  $\mu$  will be discussed shortly.) We have no good information on  $\sigma^2$ , the variance of an observation  $y_i$ , so we specify a reference prior on  $\sigma^2$  that is independent of  $\mu$ . The reference prior we use is a gamma distribution on  $\tau \equiv 1/\sigma^2$  with very small parameters. In particular, we used

$$\tau \sim \text{Gamma}(0.001, 0.001).$$

With this prior, computer simulations gave the posterior median of  $\mu$  as  $\tilde{\mu} = 4.69\%$ , that is,  $\Pr(\mu \leq 4.69 | y_1, \dots, y_{12}) = 0.5$ . For  $\sigma$  the median is  $\tilde{\sigma} = 0.64\%$ . Similarly, the 95% probability interval for  $\mu$  is from 4.49% to 4.90%. For  $\sigma$  the 95% probability interval is (0.44%, 1.03%). We are using some pretty substantial prior information. The precision of the prior on  $\mu$  is  $1/0.0163$ , about twice the estimated precision of the sample mean  $n/s^2$ , so the middle of the posterior distribution is being raised substantially from the sample mean  $\bar{y} = 4.568$  towards the prior mean 4.75.

To find the probability that zinc needs to be added to the metal before casting, let  $y_{13}$  be the zinc score of a future batch. We want to know the chances that  $y_{13} \leq 4.4$ . To evaluate this, we assume

$$y_{13} | \mu, \sigma \sim N(\mu, \sigma^2)$$

independent of  $y_1, \dots, y_{12}$  (given  $\mu$  and  $\sigma$ ). Our new parameter of interest is  $\gamma \equiv \Pr(y_{13} \leq 4.4 | \mu, \sigma)$ . If our goal is to make a prediction about whether the next outcome will be less than 4.4, we simply obtain the predictive probability  $\Pr(y_{13} \leq 4.4 | y_1, \dots, y_{12})$ . This value turns out to be 0.33. It is also the posterior mean of  $\gamma$ . So we are 1/3 sure that the outcome will be less than 4.4 based on these data and our prior input. If we are also interested in making inferences about the proportion of future outcomes that will be less than 4.4, namely the parameter  $\gamma$ , we can obtain a 95% PI in the usual way, that is, (0.20, 0.45). Note that  $\gamma = \Phi((4.4 - \mu)/\sqrt{1/\tau})$  where  $\Phi(\cdot)$  is the cdf of a  $N(0, 1)$ .

**EXERCISE 5.23.** As previously mentioned, with  $N(\mu, 1/\tau)$  data, the SIR prior is  $p(\mu, \tau) = p(\mu)p(\tau) \propto 1/\tau$ . This prior is often referred to as a Jeffreys' prior despite the fact that it is not the joint Jeffreys prior (see Exercise 4.15). In WinBUGS, we often approximate this prior with  $\mu \sim dflat()$  or  $\mu \sim N(0, 10^6)$  and  $\log(\tau) \sim dflat()$  or  $\tau \sim \text{Gamma}(0.001, 0.001)$ . (a) Perform your own analysis of the zinc data using the following WinBUGS code. With `dflat()`, you cannot use the "gen inits," you will get an error message. (b) Revise the code and reproduce the results given above where the informative prior was used for  $\mu$ . Compare results based on the two priors.

```
model{
  for(i in 1:n){ y[i] ~ dnorm(mu, tau) }
  mu ~ dflat()
  tau ~ dgamma(c,d)
  sigma <- 1/sqrt(tau)
  gamma <- phi((4.4-mu)/sqrt(1/tau))
  prob <- step(4.4 - y[13])
}
list(y=c(4.20,4.36,4.11,3.96,5.63,4.50,
       5.64,4.38,4.45,3.67,5.26,4.66,NA),
     c=0.001, d=0.001,n=13)
list(mu=0,tau =1)
```

In this exercise we are generating a predictive value by treating  $y_{13}$  as missing data. (NA stands for "not available.")

Typically, to specify an independence prior we assume  $\mu \sim N(a, 1/b)$  and that either  $\tau \sim \text{Gamma}(c, d)$  or  $\sigma \sim \text{Gamma}(e, f)$ . We must elicit information that will allow us to identify the four hyperparameters. The mean  $\mu$  is easy to think about, but the precision  $\tau$  and the standard deviation  $\sigma$  are not. We seek to infer a distribution on  $\tau$  or  $\sigma$  from parameters that are easier to think about. The precision of the data distribution is related to percentiles; percentiles of the sampling distribution are much easier to think about. Thus we induce a distribution on  $\tau$  or  $\sigma$  from information elicited about percentiles. We begin with the prior on  $\mu$ .

If we are modeling exam scores with a normal distribution, the instructor may have lots of prior data/information about long-run average exam scores for a given subject. If we are modeling wheat yields on Eastern Washington farms, there will be considerable past experience on the long-run

average wheat yield over time. The mean  $\mu$  should be easy to think about directly. Typically, we begin by eliciting an expert's best guess for  $\mu$  and set that equal to the prior mean (median and mode)  $a$ . For analyzing exam scores by a teacher, the teacher may guess that the long-term average of all scores is  $a = 65$  percent.

To get the prior variance on  $\mu$  we need to evaluate how sure we are about the guess  $a$ . If we are very sure about it, we want a large prior precision for  $\mu$ . We ask our expert to think about an upper (or lower) bound they are “virtually certain” that the mean would be below (above). In the classroom example, the instructor might be sure that the mean exam score could not be larger than 70 out of 100. We interpret this statement as  $\Pr(\mu \leq 70) = 0.95$ . The instructor is virtually certain that the mean cannot be above 70, but we add a bit of doubt so as to not be “too sure.”

Now we must find a normal distribution with a mean of  $a = 65$  and a variance (or precision) that corresponds to being 95% sure that  $\mu \leq 70$ . We know that the 95th percentile of the  $N(a, 1/b)$  distribution is  $a + 1.645\sqrt{1/b}$ . We have already specified  $a$ . We have specified the best guess for this percentile to be  $u = 70$  ( $u$  for “upper”), so we solve for  $b$ . Setting  $u = a + 1.645\sqrt{1/b}$  implies that  $(1/b) = \{(u - a)/1.645\}^2$ , or equivalently  $b = \{1.645/(u - a)\}^2$ . Substituting  $a = 65$  and  $u = 70$  gives  $b = 0.108$  and the prior on  $\mu$  for the instructor is

$$\mu \sim N(65, 1/0.108).$$

(If we had specified a lower value  $\ell$ , we would have the 5th percentile, which would identify  $b$  through  $a - 1.645\sqrt{1/b} = \ell$ .)

It is wise to try different definitions of “virtually certain,” say the 90th or 99th percentile rather than the 95th. One should examine the results and discuss them with the expert. It is also wise to elicit additional percentiles. These will rarely give exactly the same normal distribution for  $\mu$ . We need to find some compromise distribution that adequately expresses the expert's information. It may even be necessary to abandon the normal family of distributions for  $\mu$ .

**EXERCISE 5.24.** Solve for  $b$  above when the lower 5th percentile is specified by the expert. Then give both values for  $b$  (based on upper and lower values being specified), when the expert is 90% sure instead of 95% sure. For the test grades, find the  $b$ s for  $u = 70$  and  $\ell = 58$ . Find a compromise distribution for  $\mu$  that you think is reasonable. Show that  $1/b = 0.0163$  for Example 5.2.1.

It is not easy to think directly about the variance, or standard deviation, or precision but it is relatively easy to think about percentiles of the distribution of data values  $y_i$ . (Not to be confused with the percentiles of the distribution for  $\mu$  that we just used.) For example, a wheat farmer should be able to think about how many bushels per acre she would exceed 1 year out of every 10, or 9 years out of 10. Our instructor might find it easy to think about the 90th percentile of exam scores, i.e., the number  $\gamma_{0.90}$  that only 10% of students score above. It may be convenient to think about a lower percentile, say  $\gamma_{0.10}$ . Given a prior guess for the mean, and our assumption of independence of the mean and the precision/variance, information about these percentiles then gives us information about the variability. We also need to examine not only a best guess for the percentile, say  $\tilde{\gamma}_{0.10}$ , but also how accurate that guess is.

The best guess for  $\gamma_{0.90}$  provides us with a best guess for  $\tau$  and  $\sigma$ , say  $\tau_0$  and  $\sigma_0$ . We know that  $\gamma_{0.90} = \mu + 1.28\sigma$ . Generically, we write the  $\alpha$  percentile of the data as  $\gamma_\alpha$  and the  $\alpha$  percentile of a standard normal as  $z_\alpha$ . Using  $a$  and  $\tilde{\gamma}_\alpha$ , i.e., our best guesses for  $\mu$  and  $\gamma_\alpha$ , we have

$$\tilde{\gamma}_\alpha = a + z_\alpha \sigma_0 = a + z_\alpha \sqrt{1/\tau_0},$$

which gives our best guesses for the standard deviation and the precision as

$$\sigma_0 = (\tilde{\gamma}_\alpha - a)/z_\alpha \quad \tau_0 = [z_\alpha/(\tilde{\gamma}_\alpha - a)]^2.$$

For our instructor, the best guess for the mean grade  $\mu$  was  $a = 65$ . Suppose the best guess for the 90th percentile of individual test scores is  $\tilde{\gamma}_{0.90} = 85$ . That corresponds to a best guess for the standard deviation  $\sigma$  and the precision  $\tau$  of

$$\sigma_0 = 15.625; \quad \tau_0 = 0.004096,$$

respectively. But a best guess for  $\tau$  or  $\sigma$  is not enough.

We also need some idea of the uncertainty about  $\tau$  or  $\sigma$ . Using Gamma distributions to model these parameters, we set our prior guess to be the prior mode. To completely specify a Gamma distribution we elicit information that will lead to information about a percentile of this distribution. Note that if the prior mode is very close to 0, as is  $\tau_0 = 0.004096$ , it will be very difficult numerically to find a Gamma distribution with that mode and a specified small percentile (say 0.10) for it.

As good Bayesians, we presume that the parameter  $\gamma_\alpha$  is modeled with a distribution and that  $\tilde{\gamma}_\alpha$  is some measure of the center of that distribution. We now ask the expert to give us a percentile for the distribution of  $\gamma_\alpha$ . We remind the instructor that his best guess for  $\gamma_{0.9}$  was  $\tilde{\gamma}_{0.9} = 85$  and ask how much larger than 85 he believes  $\gamma_{0.9}$  could possibly be. In general, we take this upper limit, say  $\tilde{u}_\alpha$ , to be the 90th, 95th, or 99th percentile of the distribution. Of course we could work with a lower limit if convenient. Formally, if we have  $\Pr(\gamma_\alpha \leq \tilde{u}_\alpha) = 0.95$ , say, we then argue that

$$\begin{aligned} 0.95 &= \Pr(\mu + z_\alpha \sigma \leq \tilde{u}_\alpha | \mu = a) \\ &= \Pr(a + z_\alpha \sigma \leq \tilde{u}_\alpha) \\ &= \Pr(\sigma \leq (\tilde{u}_\alpha - a)/z_\alpha) \\ &= \Pr(\tau \geq \{z_\alpha/(\tilde{u}_\alpha - a)\}^2). \end{aligned}$$

Thus, the 95th percentile for  $\sigma$ , and the 5th percentile for  $\tau$  are

$$\tilde{\sigma}_{0.95} \equiv (\tilde{u}_\alpha - a)/z_\alpha; \quad \tilde{\tau}_{0.05} \equiv [z_\alpha/(\tilde{u}_\alpha - a)]^2,$$

respectively. If our instructor believes that the 90th percentile of test scores could be no higher than  $\tilde{u}_{0.90} = 91$ , then with  $a = 65$  we have

$$\tilde{\sigma}_{0.95} = 20.31; \quad \tilde{\tau}_{0.05} = 0.00242367.$$

Let's first find a  $\text{Gamma}(e, f)$  distribution for  $\sigma$  that is centered near  $\sigma_0$ , and has 95th percentile  $\tilde{\sigma}_{0.95}$ . Equate the mode of the  $\text{Gamma}(e, f)$  to  $\sigma_0$ , so

$$\sigma_0 = (e - 1)/f \quad \text{or} \quad e = 1 + \sigma_0 f. \quad (5)$$

If we can specify a value for  $f$ , we are done. We need to find  $f$  so that the  $\text{Gamma}(1 + f\sigma_0, f)$  distribution has 95th percentile  $\tilde{\sigma}_{0.95}$ . This can be accomplished by trial and error using any computer routine that finds percentiles of Gamma distributions. Just keep trying different values of  $f$  until you find one that has  $\tilde{\sigma}_{0.95}$  as its 95th percentile. For our illustration, we require a mode of 15.625 and a 95th percentile of 20.31. This occurs with a  $\text{Gamma}(41.4375, 2.588)$  distribution. Overall, our independence prior for the instructor is

$$\mu \sim N(65, 1/0.108) \quad \perp\!\!\!\perp \quad \sigma \sim \text{Gamma}(41.4375, 2.588).$$

Alternatively, we might find a  $\text{Gamma}(c, d)$  distribution for  $\tau$  that has mode

$$\tau_0 = (c - 1)/d \quad \text{or} \quad c = 1 + \tau_0 d \quad (6)$$

and 5th percentile  $\tilde{\tau}_{0.05}$ , which is again accomplished by trial and error. For our illustration, we have a mode of  $\tau_0 = 0.004096$  and 5th percentile of  $\tilde{\tau}_{0.05} = 0.00242367$ . These quantities are very close to 0 and Gamma distributions with small modes will be highly skewed. It is difficult to find a Gamma

distribution that satisfies these constraints. We recommend that information be elicited about the 5th percentile of  $\gamma_\alpha$  rather than the 95th. This leads to matching the mode and the 95th percentile of the Gamma distribution. In particular,  $\Pr(\gamma_\alpha > \tilde{\ell}_\alpha | \mu = a) = 0.95$ , which leads to  $\Pr\{\tau < [z_\alpha / (\tilde{\ell}_\alpha - a)]^2\} = 0.95$ . We find the Gamma distribution with mode  $\tau_0$  and 95th percentile  $\tilde{\tau}_{0.95} = [z_\alpha / (\tilde{\ell}_\alpha - a)]^2$ . For our example, we might be virtually certain that  $\gamma_{0.90} > 79 = \tilde{\ell}_\alpha$ , in which case, with  $a = 65$ , we get  $\tilde{\tau}_{0.95} = 0.008359$ . Trial and error leads to a  $\text{Gamma}(6.439, 1328)$  distribution. This leads to the alternative independence prior for our instructor of

$$\mu \sim N(65, 1/0.108) \quad \perp \!\!\! \perp \quad \tau \sim \text{Gamma}(6.439, 1328).$$

This distribution on  $\tau$  induces a distribution on  $\sigma$  that we can picture by simulating 10,000 variates from the  $\text{Gamma}(6.439, 1328)$  distribution, taking one over the square root of each of them, and getting a histogram. Compare this with the  $\text{Gamma}(41.4375, 2.588)$  prior on  $\sigma$  by getting another histogram and setting the histograms side by side. We did this and found that the histograms looked quite similar but the Gamma distribution for  $\sigma$  was shifted slightly to the right and more symmetric. Subject to further expert information, we believe that either prior would be satisfactory.

*To eliminate problems associated with very small numerical values, we often examine the best guesses  $\tau_0$  and  $\sigma_0$ , note which one is greater than one, and develop a Gamma prior for the corresponding parameter.*

Prior elicitation is hard work. We would not blame anyone who just quit after providing the best guesses  $\tau_0$  and  $\sigma_0$ . If working with  $\tau$ , a simple procedure specifies a Gamma prior with mode of  $\tau_0$  and a large variance. Since the variance of a  $\text{Gamma}(c, d)$  is  $c/d^2$ , we pick a small value for  $d$ , say 0.001 and find the appropriate  $c$ . Using (6), matching the mode gives  $\tau \sim \text{Gamma}(1 + \tau_0/1000, 1/1000)$  with mean  $1000 + \tau_0$  and standard deviation  $\sqrt{1000}\sqrt{1000 + \tau_0}$ . For our school example, this leads to a  $\text{Gamma}(1.000004, 0.001)$  distribution. The hope is that this will be similar to a recentered version of the Jeffreys' prior, one that will have relatively little effect on the data other than a little recentering. Alternatively, using (5) we could specify a prior on  $\sigma \sim \text{Gamma}(1 + \sigma_0/1000, 1/1000)$ . In the example,  $\sigma \sim \text{Gamma}(1.0156, 0.001)$ . Alternatively, we could place a log-normal prior on  $\tau$  so that  $\log(\tau) \sim N(\log(\tau_0), 10^6)$  with a similar result for  $\sigma$ .

The use of  $d = f = 1/1000$  in these Gamma distributions is somewhat arbitrary. If we can identify the order of magnitude of the precision or standard deviation, alternative values of  $d$  and  $f$  should be investigated. For example, when looking at test score percentages,  $\sigma$  could be 10 but it is unlikely to be 100. In that case we should consider using  $\sigma \sim \text{Gamma}(1 + \sigma_0/100, 1/100)$  or even  $\sigma \sim \text{Gamma}(1 + \sigma_0/10, 1/10)$ .

**EXERCISE 5.25.** Explain how the independence of  $\mu$  and  $\tau$  was used in the development. It was actually used twice, once explicitly without mention and the other time only implicitly.

**EXERCISE 5.26.** (a) Run R code # 1 below to verify that we obtained the correct Gamma prior for  $\sigma$  in the exam score example. (b) Run # 2 to place a comparable prior on  $\tau$ . (c) Simulate values from both priors in R or WinBUGS and create a histogram or plot for the actual Gamma prior for  $\sigma$  in the one case and the induced prior for  $\sigma$  in the second case. Compare them. (e) Modify the R code to obtain a prior on  $\mu$  and  $\sigma$  when (i) your best guess for the mean exam score is 60, (ii) you are 95% sure that the mean exam score is less than 65, (iii) your best guess for the 90th percentile of exam scores is 80, and (iv) you are 95% sure that the 90th percentile is less than 90. (d) Modify code #2 below to obtain a prior for  $\tau$  when you are 95% sure that  $\gamma_{0.90} > 70$ , and using the pertinent information in part (c).

```
#1 R Code for finding prior on sigma
alpha <- 0.90
beta <- 0.95
a <- 65 # Best guess for mu
tildegamma <- 85 # Best guess for gamma_alpha
```

```

tildeu <- 91 # Best guess percentile of gamma_alpha
zalpha <- 1.28 # qnorm(0.90,0,1)
f <- 3 # Initial value for f
# Could use a sequence of values, say f <- seq(1,50,1)
sigma0 <- (tildegamma - a)/zalpha
e <- 1 + sigma0*f
# We must find the Gamma(e,f) distribution that
# has beta-percentile = tildesigmabeta
tildesigmabeta <- (tildeu - a)/zalpha
trialq <- qgamma(beta,e,f) # Return beta-percentile for the
# selected gamma distribution
trialq      # If trialq = tildesigmabeta
tildesigmabeta # stop and pick corresponding f

#2 R Code for finding prior on tau
alpha <- 0.90
beta <- 0.95
zalpha <- 1.28
a <- 65
tildegamma <- 85
tildel <- 79
d <- 1100
tau0 = (zalpha/(tildegamma - a))^2
c = 1 + tau0*d
tildetaubeta = (zalpha/(tildel - a))^2
trialq = qgamma(beta,c,d)
trialq
tildetaubeta

```

#### 5.2.4 Some Curious Distributional Results\*

We have been considering observations

$$y_1, \dots, y_n | \mu, \sigma^2 \stackrel{iid}{\sim} N(\mu, 1/\tau)$$

with independence prior

$$\mu \sim N(a, 1/b) \quad \perp \!\!\! \perp \quad \tau \sim \text{Gamma}(c, d).$$

As mentioned earlier, this is our first real example in which the posterior distribution cannot be obtained by calculus; it requires simulation. We now look at the posterior in more detail and find some curious distributional results.

As in (4),

$$L(\mu, \tau | y) \propto \tau^{n/2} \exp \left[ -\frac{\tau}{2} (n-1)s^2 - \frac{\tau}{2} n(\bar{y} - \mu)^2 \right].$$

With the independent normal and Gamma priors on  $\mu$  and  $\tau$ , respectively, the joint posterior density becomes

$$p(\mu, \tau | y) \propto \tau^{n/2} e^{-(\tau/2)\{n(\bar{y}-\mu)^2+(n-1)s^2\}} e^{-(b/2)(\mu-a)^2} \tau^{c-1} e^{-\tau d}. \quad (7)$$

Similar to the formula for conjugate priors in Subsection 5.2.2, write

$$\hat{\mu} \equiv \hat{\mu}(\tau) \equiv \frac{n\tau}{n\tau+b}\bar{y} + \frac{b}{n\tau+b}a.$$

From Exercise 5.27 below, the posterior can be rewritten as

$$p(\mu, \tau | y) \propto e^{-\frac{1}{2}(n\tau+b)(\mu-\hat{\mu})^2} \tau^{(c+n/2)-1} e^{-(\tau/2)\{2d+(n-1)s^2+(n\tau b/n\tau+b)(\bar{y}-a)^2\}}. \quad (8)$$

No matter how hard we stare at the posterior or manipulate it, we have not been able to write it in any truly convenient form, certainly not one that would allow us to identify it as a particular distribution.

The curious results are that although there is not much we can do with the overall posterior, we can find the conditional posteriors  $\mu | \tau, y$  and  $\tau | \mu, y$  exactly. The density of  $\mu | \tau, y$  is obtained by dropping any multiplicative terms that do not involve  $\mu$ . From (8)

$$p(\mu | \tau, y) \propto \exp \left\{ -\frac{1}{2} (n\tau + b) [\mu - \hat{\mu}(\tau)]^2 \right\}.$$

This is recognizable as a  $N[\hat{\mu}, 1/(n\tau + b)]$  density. Moreover, dropping terms from (7) that do not involve  $\tau$  we similarly obtain

$$p(\tau | \mu, y) \propto \tau^{\{c+n/2\}-1} \exp(-\tau [d + \{n(\bar{y} - \mu)^2 + (n-1)s^2\}/2]),$$

which is the kernel of a  $\text{Gamma}(c+n/2, d + \{n(\bar{y} - \mu)^2 + (n-1)s^2\}/2)$  density. This is also a  $\text{Gamma}(c+n/2, d + \sum_i (y_i - \mu)^2/2)$  density, which is algebraically simpler but computationally more intense when frequently changing  $\mu$ .

We will see in Subsection 6.3.1 that these conditional distributions are precisely what we need to sample from the posterior distribution using the Gibbs sampler. To obtain a sample  $\{(\mu^r, \tau^r) : r = 1, 2, \dots, m\}$ , if we have starting values  $(\tau^1, \mu^1)$ , we sample first from  $p(\mu | \tau^1, y)$  to obtain  $\mu^2$ . Then we sample from  $p(\tau | \mu^2, y)$  to obtain  $\tau^2$ . Continuing for  $m$  pairs of steps gives the sample. Under fairly mild conditions, after an initial “burn-in” phase of  $BI$  steps, the pairs  $\{(\mu^r, \tau^r) : BI < r \leq m\}$  constitute an identically distributed, but not necessarily independent, sample from the joint density  $p(\mu, \tau | y)$ . The entire process is called a Markov chain because values generated at the  $r$ th iteration depend only on the values generated at the  $(r-1)$ st iteration. Simulation based on Markov chains, that is, MCMC simulation, is discussed in Section 6.3.

What happens if we specify independence priors on  $\mu$  and  $\sigma$  with  $\mu \sim N(a, 1/b)$  and  $\sigma \sim \text{Gamma}(e, f)$ ? The full conditional for  $\mu | \sigma, y$  is the same with  $\tau$  replaced by  $1/\sigma^2$ . The kernel of the full conditional for  $\sigma | \mu, y$  is easily written down, but it is no longer recognizable as a Gamma or any other standard distribution. While this may seem disconcerting at first, there are methods available for sampling from unrecognizable probability distributions for which the kernel of the density is known. WinBUGS incorporates such methods so, while it may run a little slower than when the conditional distribution is known, we anticipate no problems with Markov chain sampling.

**EXERCISE 5.27.** (a) Rewrite (7) as

$$p(\mu, \tau | y) \propto \exp \left\{ -\frac{1}{2} [n\tau(\mu - \bar{y})^2 + b(\mu - a)^2] \right\} \tau^{(n/2+c)-1} e^{-\frac{\tau}{2}[(n-1)s^2+2d]}$$

and use the complete the square formula of Exercise 5.22 to obtain (8). (b) Parameterize the normal model in terms of  $\sigma$  rather than  $\tau$ . Give the kernel of the full conditional for  $\sigma$ . Simplify as best you can.

### 5.2.5 Two-Sample Normal Model

If we observe independent samples from two normal populations, often the goal is to compare the two populations. For example, we might compare student satisfaction from two methods of teaching introductory statistics: Bayesian and frequentist. Or we might compare wheat yields for two

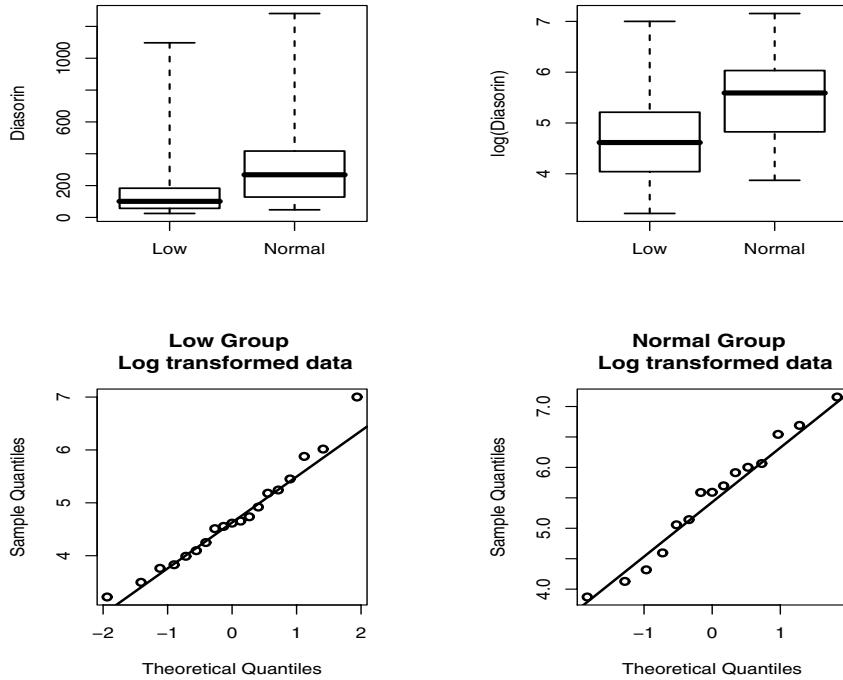


Figure 5.2: *Diasorin* data: Top left, boxplots of original data. Top right, boxplots of transformed data. Bottom, normal plots for transformed data.

kinds of fertilizer. Or perhaps we want to examine whether the average income for male computer programmers differs from that for their female counterparts. Such comparisons are greatly facilitated when the populations have the same variance.

**EXAMPLE 5.2.2.** Renal osteodystrophy is a bone disease that occurs when the kidneys fail to maintain proper levels of calcium and phosphorus in the blood. Monitoring patients with loss of kidney function for lower than normal bone turnover aids in managing the disease. A commercially available diagnostic assay, Diasorin, claims to be able to tell patients apart who have low versus normal bone turnover. A cross-section of 34 kidney patients from the bone registry at the University of Kentucky were identified as low or normal turnover by other means and then given the commercial assay to determine whether it could correctly identify them. From boxplots a normal sampling model appears untenable due to marked skewness, but boxplots and quantile plots of the log transformed data seem reasonably normal, see Figure .

In general we assume the sampling model

$$y_{11}, \dots, y_{1n_1} | \mu_1, \tau_1 \stackrel{iid}{\sim} N(\mu_1, 1/\tau_1) \quad \perp\!\!\!\perp \quad y_{21}, \dots, y_{2n_2} | \mu_2, \tau_2 \stackrel{iid}{\sim} N(\mu_2, 1/\tau_2).$$

We typically assume independent priors for the two populations

$$\mu_1, \tau_1 \quad \perp\!\!\!\perp \quad \mu_2, \tau_2.$$

We can use any of the one-sample techniques—reference priors, conjugate priors, or independence priors—to obtain priors for the parameters of each population.

Inequality of variances (*heteroscedasticity*) is a problem. Unlike frequentist inference, it poses no technical problems to Bayesians. It is straightforward to obtain posterior inferences by simulation after incorporating unequal variances into a Bayesian model. But for Bayesians and frequentists alike, the real issue is one of interpretation. What does it mean if  $\mu_2$  is larger than  $\mu_1$  but the standard deviation for population 1 is much larger than for population 2? Figure 5.3 displays two such densities. If larger responses are preferable, the average response in population 2 is preferable to that for population 1. But there is a sizeable proportion of individuals from population 1 who have higher responses than those from population two. For example, we are often interested in whether some cut-off value is exceeded: systolic blood pressure over 140, diastolic over 90, cholesterol score over 240. In Figure 5.3, population 1 has a much higher percentage of people with scores above 7 than population 2, even though  $\mu_1 < \mu_2$ . Of course if  $\mu_2$  is so much larger than  $\mu_1$  that the difference dwarfs the standard deviations, virtually all of population 2 will be larger than population 1. In any particular application, it is not obvious from the means alone which population is preferable. With equal variances, the population means tell much more of the story. *We find that plotting the predictive distribution is a most useful tool, especially when analyzing normal data with unequal variances.*

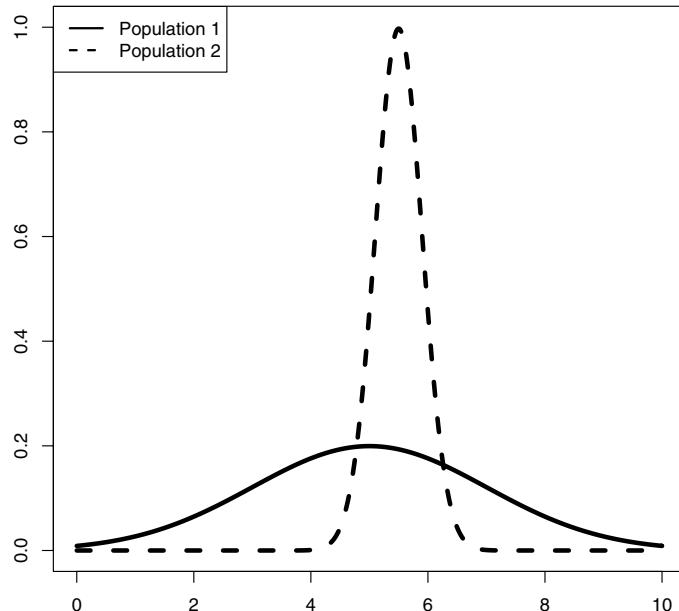


Figure 5.3: Two normal distributions:  $N(5, 2^2)$  and  $N(5.5, .4^2)$ .

**EXAMPLE 5.2.2 CONTINUED.** We elicit independence priors for each population. Although we analyze data on the log scale, we elicit priors on the original scale. To get a prior on a mean  $\mu$ , we elicit information on the median  $m = e^\mu$  of the sampling distribution. We specify a best guess  $\tilde{m}$  for the median, and a percentile  $\tilde{u}$  for which we are, say, 95% sure that the median is below (or above). Percentiles are invariant under the log transformation, thus  $\Pr[\mu \leq \log(\tilde{m})] = 0.5$  and  $\Pr[\mu \leq \log(\tilde{u})] = 0.95$ . If we model  $\mu \sim N(a, b)$ , then  $a = \log(\tilde{m})$  and  $[\log(\tilde{u}) - \log(\tilde{m})]/\sqrt{b} = 1.645$ , so  $b = [\log(\tilde{u}) - \log(\tilde{m})]^2/1.645^2$ .

Prior information for the Diasorin analysis was elicited from Dr. Johann Herberth MD, PhD, MPH (Division of Nephrology, Bone and Mineral Metabolism; University of Kentucky). The

Table 5.3: Posterior summaries for informative prior.

Parameter	mean	sd	2.5%	50%	97.5%
$\mu_1$	4.860	0.052	4.757	4.860	4.961
$\mu_2$	5.395	0.051	5.294	5.395	5.496
$\mu_2 - \mu_1$	0.536	0.073	0.392	0.536	0.679
$\tau_1$	1.275	0.394	0.625	1.231	2.161
$\tau_2$	1.285	0.439	0.576	1.236	2.274
$\tau_2/\tau_1$	1.114	0.555	0.387	1.003	2.495

median for the low bone turnover group ( $n_1 = 19$ ) was believed by Dr. Herberth to be  $\tilde{m}_L \equiv \tilde{m}_1 = 130$  and he was 95% sure that this median was less than  $\tilde{u}_L \equiv \tilde{u}_1 = 142$  in this patient population. These give  $\mu_L \equiv \mu_1 \sim N(4.87, 0.00288)$ . The corresponding values elicited from Dr. Herberth for the normal bone turnover group ( $n_2 = 15$ ) were  $\tilde{m}_N \equiv \tilde{m}_2 = 220$  and  $\tilde{u}_N \equiv \tilde{u}_2 = 240$ , respectively, leading to  $\mu_N \equiv \mu_2 \sim N(5.39, 0.00280)$ .

The priors on the  $\tau$ s are also obtained as for the one-sample normal case. We elicit information from our expert about the 90th (or some other) percentile of the sampling distribution  $\gamma_{0.9}$ . The log of this is  $\mu + 1.645\sqrt{1/\tau}$ . The elicitation is now conditional on the best guess for  $\mu$  being  $\log(\tilde{m})$ . Suppose our best guess for the 90th percentile is  $\tilde{\gamma}_{0.9}$ . Then our best guess for  $\tau$ , say  $\tau_0$ , conditional on our best guess for  $\mu$ , is obtained by solving  $\log(\tilde{\gamma}_{0.9}) = \log(\tilde{m}) + 1.645\sqrt{1/\tau_0}$ . We obtain  $\log(\tilde{\gamma}_{0.9}/\tilde{m}) = 1.645\sqrt{1/\tau_0}$  or  $\tau_0 = 1.645^2 / \{\log(\tilde{\gamma}_{0.9}/\tilde{m})\}^2$ . Since we assume a Gamma( $c, d$ ) prior for  $\tau$ , set  $(c-1)/d = \tau_0$  or equivalently  $c = 1 + \tau_0 d$ . We can proceed with eliciting an upper limit on  $\gamma_{0.9}$  but often the expert wants to stop, in which case we introduce the same large variability as in a proper gamma reference prior by picking a small value for  $d$ .

Dr. Herberth provided his best guess for the 90th percentile of Diasorin values in the low ( $\tilde{\gamma}_{0.90,1} = 170$ ) and normal ( $\tilde{\gamma}_{0.90,2} = 280$ ) bone turnover groups. We consider gamma priors with modes  $\tau_{0,1} = 37.60$  and  $\tau_{0,2} = 46.53$ . First consider  $d = 0.001$ :  $\tau_1 \sim \text{Gamma}(1.0376, 0.001)$  and  $\tau_2 \sim \text{Gamma}(1.04653, 0.001)$ . One concern with this choice is that the means of the  $\tau$ s are on the order of 1000, so these priors give lots of vaguely specified probability to the right of the mode. In fact, they give 97% and 96% prior probabilities to being greater than their modes, respectively. If we let  $b = 0.01$ , these probabilities reduce to 0.7 and 0.65, and if we let  $b = 0.1$ , they are 0.025 and 0.01, respectively. We still used  $b = 0.001$  but did a sensitivity analysis. Another choice here would be to simply use  $p(\tau_j) = 1/\tau_j$ , or its approximation, a Gamma(0.001, 0.001) for each. These have a mode of 0, a mean of 1, and a variance of 1000. Our sensitivity analysis also involves selecting normal priors for the  $\mu$ s but with larger variances.

A common summary of results for two normal samples is given in Table 5.3 computed with our informative prior. Although the median of the normal group is larger than that of the low group ( $\Pr[\mu_2 > \mu_1 | y] \doteq 1$ ), the predictive probability that a new outcome from the normal group exceeds an outcome for the low group is only 0.66. To address allocation issues more carefully, we used  $5.1 \doteq (\hat{\mu}_1 + \hat{\mu}_2)/2$  as a cutoff to decide who should be assigned to the low group and who should be assigned to the normal group based on their log Diasorin scores. The predictive probability of someone from the normal group getting above this value is about 63%. The predictive probability of someone from the low group scoring below 5.1 is 61%. So, with this cutoff value, whether you are normal or low, there is nearly a 40% chance that you will get misclassified.

Figure 5.4 shows predictive densities, using Dr. Herberth's prior, for a future *log Diasorin value* from the low and normal groups. Note the similarity of the distributional shapes, which is due to the similarity of the precisions. With similar precisions, it becomes clear that the “normal” group tends to have higher scores and that the means characterize the differences between the two distributions. If the variances were not similar, the difference in means would not be nearly so meaningful. To illustrate this, Figure 5.5 gives the predictive distributions for untransformed Diasorin scores. These densities are more difficult to interpret relative to one another. In particular, the difference in the

means of the predictive distributions by itself does not seem like a particularly good measure of how the two distributions differ. Even with unequal variances, it is probably easier to evaluate the differences between groups on the log scale.

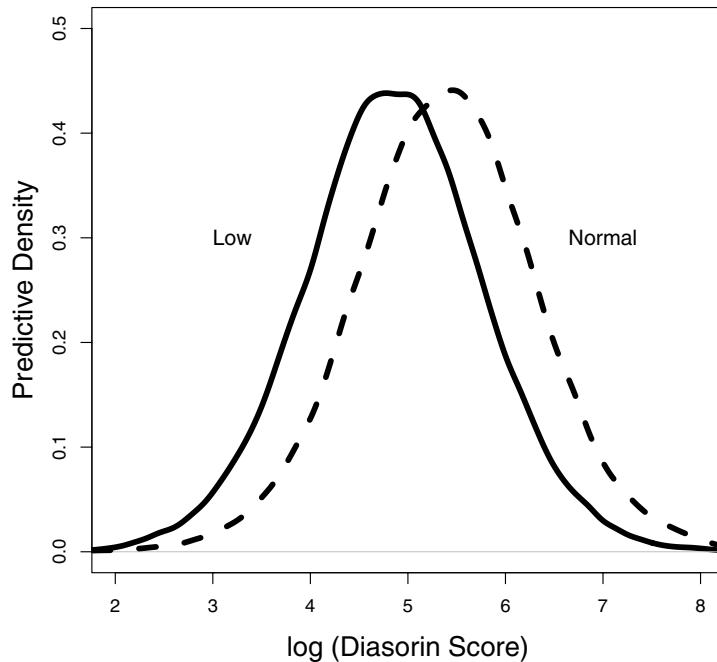


Figure 5.4: *Diasorin data: Predictive densities on log scale.*

We commonly need to transform data before a normal analysis. We now illustrate drawing conclusions on the original scale. Table 5.4a gives posterior summaries for the two groups based on our informative prior. The predictive distributions are so skewed that means and standard deviations provide little useful information, so we do not report them. We focus on medians and probability intervals. The posterior median of  $e^{\mu_L}$ , the Diasorin value among low bone turnover patients was  $e^{\tilde{\mu}_L} = 129$  with 95% PI (116.4, 142.8) and for patients with normal bone turnover the median Diasorin score  $e^{\mu_N}$  has posterior median 220.6 with 95% PI (199.1, 243.6). The relative median comparing the Diasorin values of normal to low bone turnover patients is  $e^{\mu_N - \mu_L}$  with posterior median 1.71 and 95% PI (1.48, 1.97). Prediction intervals for new values from both groups are also given.

We performed a sensitivity analysis. First, we set  $d = 0.01$  in the priors for  $\tau_j$ . There was virtually no effect on the estimates in the informative analysis. We then considered priors with very large variances for the  $\mu$ s, but which had the same means as the informative prior. Those results are given in Table 5.4b. The results change noticeably, but not dramatically. The prior on  $\mu$  obviously had an effect on the analysis. Finally, in Table 5.4c, we present results based on the proper reference priors given earlier. For the most part, there are small changes from the analysis in (b), except for the upper end of the prediction intervals, which are considerably higher with this prior. This occurs because of the expert information that remains in the prior of part (b). A little prior information has given more precision in the predictions, although the intervals are very wide.

There are substantial areas of agreement in the various analyses presented in Table 5.4. The median for the low group is probably in the low 100s. The median for the normal group is probably

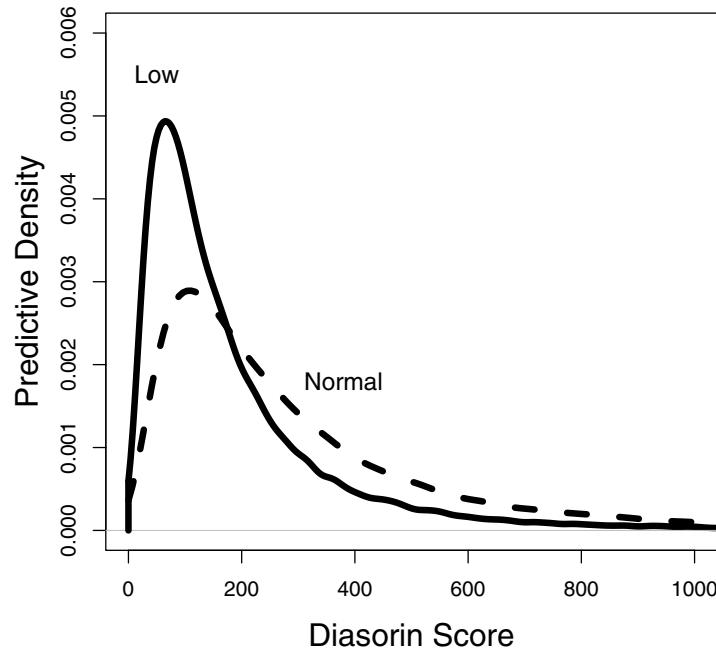


Figure 5.5: *Diasorin data: Predictive densities on original scale.*

in the mid to low 200s. There is clearly a difference in the groups. To make more specific claims than these probably requires more data. Although there is clearly a difference in the scores between the two groups, just from looking at Figures 5.4 and 5.5 it is by no means clear that the difference is sufficient to be clinically useful in diagnosing low bone turnover.

The traditional approach to comparing two different normal populations is to look at  $\mu_1 - \mu_2$ . However, when we allow for distinct variances (precisions) in the two populations, a simple comparison of means can be meaningless. Differences in variances can create huge changes in the practical effects associated with a difference in the means. The problem of making inferences on means when the variances differ has a long and controversial history with little agreement even today. We believe that the best practical method for dealing with this problem is by examining the predictive distributions and carefully interpreting them. However, if we can assume  $\tau_1 = \tau_2 = \tau$ , not only is there is a nice analytical Bayesian solution for finding the posterior distribution of  $\mu_1 - \mu_2$  using conjugate priors, but there is no controversy about  $\mu_1 - \mu_2$  being the correct “effect measure” to examine.

**EXERCISE 5.28.** When  $\log(y) \equiv w \sim \text{Normal}(\mu, 1/\tau)$ ,  $y$  is said to have a log-normal distribution. Log-normal distributions are often used in analyzing time to event data. Use Proposition B.4 to derive the density for  $y$ .

**EXERCISE 5.29.** Simulate two samples from distinct normal distributions with different sample sizes, means, and precisions. Using independent proper reference priors as discussed for the single-sample normal case, employ the following WinBUGS code to make inferences about both means and precisions. Try this with a number of different parameter settings for the simulated data and explore how easy or difficult it is to detect a difference in means.

Table 5.4: Posterior summaries on original scale for Diasorin data.

(a) Informative Prior					
Inference	mean	sd	2.50%	median	97.50%
Pred Dens L	—	—	20.05	128.0	821.7
Pred Dens N	—	—	34.43	219.9	1416.0
Med <sub>L</sub>	129.1	6.738	116.4	129.0	142.8
Med <sub>N</sub>	220.6	11.35	199.1	220.3	243.6
Med <sub>N</sub> /Med <sub>L</sub>	1.713	0.1257	1.479	1.709	1.972
(b) Large Variance for $\mu$					
	mean	sd	2.50%	median	97.50%
Pred Dens L	—	—	15.7	110.2	768.5
Pred Dens N	—	—	34.06	241.9	1752.0
Med <sub>L</sub>	113.2	24.95	72.09	110.7	169.4
Med <sub>N</sub>	249.7	63.52	147.8	242.3	396.1
Med <sub>N</sub> /Med <sub>L</sub>	2.311	0.7883	1.141	2.188	4.183
(c) Proper Reference Prior					
	mean	sd	2.50%	median	97.50%
Pred Dens L	—	—	13.26	110.5	921.5
Pred Dens N	—	—	28.53	242.3	2069.0
Med <sub>L</sub>	113.4	26.48	70.2	110.5	173.8
Med <sub>N</sub>	251.3	70.09	141.6	242.3	414.2
Med <sub>N</sub> /Med <sub>L</sub>	2.336	0.8637	1.09	2.193	4.418

```

model{
  for(i in 1:n[1]) { y[i] ~ dnorm(mu[1], tau[1]) }
  for(j in 1:n[2]) { x[j] ~ dnorm(mu[2], tau[2]) }
  for(r in 1:2){
    mu[r] ~ dnorm(a[r], b[r])
    tau[r] ~ dgamma(c[r], d[r])
    sigma[r] <- sqrt(1/tau[r])
  }
  meandiff <- mu[1] - mu[2]
  sdratio <- sigma[1]/sigma[2]
  prob[1] <- step(meandiff) # Gives Pr(meandiff >0|data)
  prob[2] <- step(sdratio -1) # Gives Pr(sdratio >1|data)
}
list(y = c(.,.,.,.), x = c(.,.,.,.), n = c(.,.,.,.),
a = c(.,.,.), b = c(.,.,.), c = c(.,.,.), d = c(.,.,.))
list(mu = c(0,0), tau = c(1,1))

```

To simulate normals in R, use the command `rnorm(n, mu, sqrt(1/tau))`, where `n` is the sample size, `mu` is the mean, and `tau` is the precision desired.

**EXERCISE 5.30.** The WinBUGS code below was used to analyze the Diasorin data. Perform an additional sensitivity analysis by selecting a range of priors and looking at the impact of changing the priors on the final inferences. Rather than taking logs of the data, we specify log-normal distributions, but as established in Exercise 4.8, this does not matter. One set of informative priors has been “commented out.”

```

model{
  for(i in 1:n[1]){ low[i] ~ dlnorm(mu[1], tau[1]) }

```

```

for(i in 1:n[2]){
  normal[i] ~ dlnorm(mu[2], tau[2])
}
# mu[1] ~ dnorm(4.87, 347.12)
# mu[2] ~ dnorm(5.39, 357.14)
mu[1] ~ dnorm(0,0.00001)
mu[2] ~ dnorm(0,0.00001)
# tau[1] ~ dgamma(1.0376,0.001)
# tau[2] ~ dgamma(1.04653,0.001)
tau[1] ~ dgamma(0.001,0.001)
tau[2] ~ dgamma(0.001,0.001)
med[1] <- exp(mu[1])
med[2] <- exp(mu[2])
rmed <- med[2] / med[1]
test[1] <- step(med[2]-med[1])
test[2] <- step(Nf-Lf)
Lf ~ dlnorm(mu[1], tau[1])
Nf ~ dlnorm(mu[2], tau[2])
dmu <- mu[2]-mu[1]
rtau <- tau[2]/tau[1]
}
list(n=c(19,15),
      low = c(91,46,95,60,33,410,105,43,189,1097,
              54,178,114,137,233,101,25,70,357),
      normal = c(370,267,99,157,75,1281,48,298,
                268,62,804,430,171,694,404))
list(mu=c(0,0), tau=c(1,1), Lf=50, Nf=50)

```

### 5.3 Inference for Rates

In this section, we discuss the Poisson distribution for modeling one and two samples. The three most commonly used distributions for modeling data are the binomial, normal, and Poisson. The binomial is useful because it is simple: the number of successes in a fixed number of independent trials (with common probability of success). The normal is useful because many measurements involve a large number of small errors added together, so the central limit theorem suggests that the measurements can be approximated by a normal distribution. The Poisson is also relatively simple.

Poisson distributions are used to model counts. Examples include (i) the number of phone calls in a day, (ii) the number of raisins in a cookie, (iii) the number of dry oil wells in Texas, and (iv) the number of blips on a Geiger counter within 10 minutes. With binomial data we know that a count cannot exceed the fixed number of trials  $n$ . A Poisson variate has no obvious upper limit but it is usually associated with a time or location limit, e.g., a day, a cookie, a state, 10 minutes.

We agree with the ancient Greeks that the youth of today are dissolute miscreants and that civilization cannot possibly last another 30 years. So let's think about the number of BBs that are projected onto a poorly paid school teacher's poorly constructed table. Each BB leaves a mark on the table. First we assume that the BBs are projected independently, e.g., nobody has glued two BBs together so they will leave marks right next to each other. The marks are small and the table is large so we want a system for counting them. We put a grid over the table and count how many grid squares have a mark in them. The number of grid squares with marks in them will actually have a binomial distribution. But suppose we make the grid finer and finer. As the squares get smaller in area, the probability of seeing a mark in any square gets smaller but the number of squares gets larger. As the squares get smaller, there would be virtually no chance of two BBs hitting in the same square, so counting the number of squares with marks will be the same as counting the number marks. Let  $\theta$  be the expected value of the number of squares with marks in them.  $\theta$  is the expected value of a binomial but it is also the expected number of marks on the table, and the expected value

of the number of marks on the table should have nothing to do with the size of the grid that we arbitrarily put over the table. The Poisson density is derived from the binomial density by letting the number of grid squares go to infinity but requiring the expected value of the number of marks to remain the same. We say a random variable  $y$  has a  $\text{Pois}(\theta)$  distribution provided

$$f(y|\theta) = \frac{\theta^y e^{-\theta}}{y!}, \quad y = 0, 1, 2, \dots$$

The mean and variance are both  $\theta$ .

This definition of a Poisson works fine for looking at one table. If we want to estimate  $\theta$  we might want a random sample of tables, but they better all be the same size. If they are not, we need to reparameterize the distribution to adjust for table size. Let  $M_i$  be the area of table  $i$ . The appropriate model has independent observations with

$$y_i|\theta \sim \text{Pois}(\theta M_i).$$

Now  $\theta$  becomes the rate at which BBs hit a “standard” table with area 1, while  $\theta M_i$  is the mean number of marks for table  $i$ . If the tables were all of the same size, we would not bother with this; we would take  $M_i = 1$  for all  $i$  and let  $\theta$  be the mean marking rate.

More generally, a random variable  $y$  has a Poisson distribution when it gives the number of “events” in a fixed window (period of time or region in space) where

- (a) The probability of more than one “event” in a small sub-window of time (or region in space) is very small. So events occur “rarely” in this sense. We don’t get two phone calls in exactly the same instant of time.
- (b) The probability of exactly one “event” in a small sub-window of time (or region in space) is proportional to  $\theta$  times the size of the time period (or region).
- (c) The number of events occurring in non-overlapping sub-windows of time (or regions in space) are independent. This is called “independent increments.”
- (d) The distribution of the number of events in any sub-window of time (or region in space) only depends on the parameter  $\theta$ . No matter which period of time or region in space might be considered, the rate of occurrence of events is  $\theta$  per unit time (or space). This assumption is termed “stationarity.”

The next subsections consider one Poisson sample, informative priors, reference priors, and two-sample Poisson sampling. Extensions to more complex Poisson models are developed in Chapter 11.

### 5.3.1 One-Sample Poisson Data

In Section 1.3 we discussed Ache armadillo kills in Paraguay. Specifically, our data involved the number of armadillos killed in one day by each of  $n = 38$  Ache men. We modeled the data as Poisson.

Consider a sample of counts  $y = (y_1, \dots, y_n)'$  modeled as

$$y_1, \dots, y_n | \theta \stackrel{iid}{\sim} \text{Pois}(\theta).$$

The likelihood is

$$L(\theta|y) \propto \prod_{i=1}^n \theta^{y_i} e^{-\theta} = \theta^{n\bar{y}} e^{-n\theta}.$$

The conjugate prior for  $\theta$  is  $\text{Gamma}(a, b)$  with density

$$p(\theta) \propto \theta^{a-1} e^{-b\theta}.$$

The posterior density becomes

$$\begin{aligned} p(\theta|y) &\propto L(\theta|y)p(\theta) \\ &= \left(\theta^{n\bar{y}_.} e^{-n\theta}\right) \left(\theta^{a-1} e^{-b\theta}\right) \\ &= \theta^{a+n\bar{y}_.-1} e^{-(n+b)\theta}, \end{aligned}$$

so

$$\theta|y \sim \text{Gamma}(a+n\bar{y}_., b+n).$$

The gamma prior is a DAP having  $b$  prior observations with a total of  $a$  counts.

The prior and posterior have modes:

$$\frac{a-1}{b} \quad \text{and} \quad \frac{a+n\bar{y}_.-1}{b+n},$$

respectively. The posterior mean is

$$\hat{\theta} \equiv E(\theta|y) = \frac{a+n\bar{y}_.}{b+n} = \left(\frac{b}{b+n}\right) \left(\frac{a}{b}\right) + \left(\frac{n}{b+n}\right) \bar{y}_..$$

As with many DAPs, the last form writes the posterior mean as a weighted average of the sample mean  $\bar{y}_.$  and the prior mean  $a/b.$  The weight on the sample mean is large if the sample size  $n$  is large relative to the prior sample size,  $b,$  and vice versa. The posterior standard deviation is  $sd(\theta|y) = [\hat{\theta}/(b+n)]^{1/2}.$

Without computers, the Gamma distribution is hard to use. In olden days, if the first parameter was large, the Gamma distribution would be approximated by a normal distribution. If  $a+n\bar{y}_.$  is large, then

$$\theta|y \sim N\left(\hat{\theta}, \frac{\hat{\theta}}{b+n}\right).$$

**EXAMPLE 5.3.1. Ache Armadillo Hunting.** Let  $y_i$  be the number of armadillos killed by the  $i$ th man on a given day. We assume

$$y_1, \dots, y_n | \theta \stackrel{iid}{\sim} \text{Pois}(\theta)$$

and refer to  $\theta$  as the *kill rate*. Eliciting the prior

$$\theta \sim \text{Gamma}(1.11, 1.61)$$

will be discussed in the next subsection. The data give  $n = 38,$   $\sum_{i=1}^{38} y_i = 10,$  so with  $a = 1.11,$  and  $b = 1.61$  the posterior is

$$\theta|y_1, \dots, y_{38} \sim \text{Gamma}(11.11, 39.61).$$

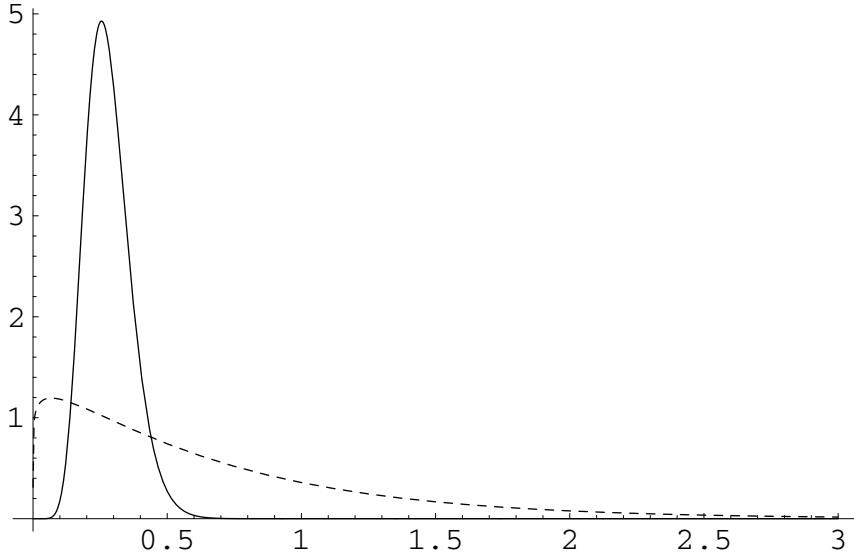
See Exercise 5.31 for the raw data.

Table 5.5 summarizes some results. After incorporating the data, with probability 0.95 the kill rate is between 0.14 and 0.47 armadillos per day, or one armadillo killed per 2  $\doteq 1/0.47$  to 7  $\doteq 1/0.14$  days. Although we know the exact posterior distribution, we need specific computer routines to find probability intervals. Or we could simulate the posterior to get probability intervals.

In this example the posterior focuses more tightly on smaller values of  $\theta$  than the prior (see Figure 5.6). The expert's prior encompasses a wide range of plausible values for  $\theta$  that are roughly centered at 0.5 and the posterior 95% probability interval falls well within the middle 95% of the prior. The data indicate that the mean number of kills is less than the expert's best guess. In fact, the posterior 95% probability interval for  $\theta$  does not contain 0.5, so we could reasonably reject the value 0.5 as implausible. The posterior is much more informative than the prior.

Table 5.5: Ache hunting: Prior and posterior medians and 95% probability intervals.

	Prior median (95% PI)	Posterior median (95% PI)
Exact	0.497 (0.024, 2.433)	0.272 (0.140, 0.468)
WinBUGS	0.498 (0.023, 2.433)	0.271 (0.140, 0.469)

Figure 5.6: Prior (dashed) and posterior (solid) distributions for kill rate  $\theta$ .

**EXERCISE 5.31.** Assume Poisson data, a  $\text{Gamma}(1,1)$  prior on  $\theta$ , and a sample of size 2 with  $y_1 = 5, y_2 = 10$ . Write and run WinBUGS code to estimate  $\theta$  and the probability of no kills,  $e^{-\theta}$ . Then obtain the normal approximation to the posterior and investigate whether it provides a reasonable approximation. Compare 95% probability intervals, posterior means, modes, and standard deviations. Modify the WinBUGS code given below, which provides results for the Ache data.

```
model{
  for(i in 1:38){ kills[i] ~ dpois(theta) }
  theta ~ dgamma(1.11,1.61)
  prob <- exp(-theta)
}
```

The data are listed in the vector kills:

```
list(kills=c(2,0,0,1,0,0,1,0,0,0,0,0,0,1,0,0,0,0,0,0,1,1,1,0,
          0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,1,0))
```

### 5.3.2 Informative Priors

Typically, we construct an informative prior for  $\theta$  by specifying a best guess and a percentile. For example, if we were modeling the number of phone calls received each day by my daughter Pandora,  $\theta$  is the mean number received. On any given day, there is variability about this mean. My best guess is that the mean is 10. I have a lot of uncertainty about the best guess of 10 but I am 95% sure that the average number of phone calls per day is fewer than 30. We want to specify a distribution that satisfies these constraints.

The mode of the  $\text{Gamma}(a, b)$  distribution is

$$\frac{a-1}{b}, \quad a \geq 1.$$

If  $a < 1$ , the mode is 0. If our best guess for  $\theta$  is  $\theta_0$ , set  $\theta_0$  equal to the mode and solve to get

$$a(b) = 1 + b\theta_0.$$

If  $b$  were known, the formula returns a value for  $a$  that gives the correct mode. In our phone example,  $a(b) = 1 + 10b$ .

In the phone example we also know

$$0.95 = \int_0^{30} \text{Gamma}(\theta|a, b)d\theta = \int_0^{30} \text{Gamma}(\theta|1+10b, b)d\theta,$$

where  $\text{Gamma}(\cdot|a, b)$  is a Gamma density. If we have a computer routine available that gives 95th percentiles of a Gamma distribution, we can find the values of  $a$  and  $b$  by trial and error. Keep trying values of  $b$  until the 95th percentile of the  $\text{Gamma}(\theta|1+10b, b)$  distribution is 30, see Exercise 5.32.

More generally, we select an upper percentile  $\alpha$  and elicit a value  $c$  such that

$$\alpha = \int_0^c \text{Gamma}(\theta|a, b)d\theta.$$

We then seek a value of  $b$  so that  $c$  is the  $\alpha$  percentile of a  $\text{Gamma}(a(b), b)$  distribution.

**EXERCISE 5.32.** Do a simple search to find  $b$  for our example with  $\theta_0 = 10$ ,  $c = 30$ , and  $\alpha = 0.95$ . Set up a grid of  $b$  values and compute the corresponding  $a(b)$  values and the 95th percentile of the  $\text{Gamma}(a(b), b)$ . Search for a percentile that agrees with  $c$ . Using the R language, commands are

```
mode <- 10
b <- 1:100
a <- 1+(mode * b)
qgamma(0.95,a,b)
```

In this case, the quantiles don't reach 30 (they range from 10 to 17), indicating that we have missed the appropriate value of  $b$ . We repeat the process using

```
mode <- 10
b <- 1:100/100
a <- 1+(mode*b)
round(qgamma(0.95,a,b),1)
```

The last command simultaneously finds the quantiles and rounds them to 1 decimal place yielding the following subset of R output

```
[1] 318.7 168.6 118.5 ... 50.9 47.4 44.6 42.2
[13] 40.2 38.4 36.9 ... 30.7 30.0 29.3 28.7
[25] 28.1 27.6
```

The appropriate percentile 30 is in the 22nd spot, so the appropriate value of  $b$  is in the 22nd spot of the  $b$  vector, which is

```
b[22]
[1] 0.22
```

The value of  $a$  is therefore  $a = 1 + (0.22)10 = 3.2$ . The prior that reflects our expert's beliefs is  $\theta \sim \text{Gamma}(3.2, 0.22)$ . Repeat this process with  $\theta_0 = 15$ .

**EXERCISE 5.33.** Just in case you forgot where we showed you how to do this, differentiate the log of the  $\text{Gamma}(a, b)$  density and set it equal to 0 to solve for the mode of the distribution.

A slightly different method was used to determine the prior for the armadillo hunting.

**EXAMPLE 5.3.1 CONTINUED.** Dr. Garnett McMillan, an expert on Ache hunting practices, believes that Ache men typically kill an armadillo every other day and thus provides a best guess for  $\theta$  of 0.5 armadillos, which we took to be the *median* of the prior distribution. Dr. McMillan is 95% sure that the kill rate is no greater than 2 armadillos. With

$$\theta \sim \text{Gamma}(a, b),$$

we solve the simultaneous equations

$$\Pr(\theta \leq 0.5|a, b) = 0.50; \quad P(\theta \leq 2|a, b) = 0.95$$

for  $a$  and  $b$  yielding  $a = 1.11$  and  $b = 1.61$ . Finding  $a$  and  $b$  is a somewhat more complicated process of trial and error than when using the mode, but it is easy to check that our prior

$$\theta \sim \text{Gamma}(1.11, 1.61)$$

matches the defining conditions (up to round-off error).

Another approach to prior construction makes use of the normal approximation. It has the advantage that the prior can be determined exactly, without a process of trial and error. With a  $\text{Gamma}(a, b)$  prior, if we anticipate that the prior will be approximately normal in shape, we could think of a  $N(a/b, a/b^2)$  prior instead. For this to work, the value of  $a$  must be moderately large and the mean and mode of the prior should be nearly the same. We won't actually know whether that is true until we calculate the values of  $a$  and  $b$ .

The method of selection is the same as before. We pick a best guess for  $\theta$  and a percentile. If our best guess for  $\theta$  is  $\theta_0$ , we set  $\theta_0 = a/b$  (since the mode should approximately equal the mean in this case). If we let  $c$  be the 95th percentile, then

$$c = \theta_0 + 1.645\sqrt{\frac{\theta_0}{b}} = \theta_0 + 1.645\frac{\theta_0}{\sqrt{a}}.$$

This equation can be solved for  $a$ , and then the value of  $b$  is obtained from the relationship  $\theta_0 = a/b$ . For example, if we specify  $\theta_0 = 50$  and  $c = 60$  we get  $60 = 50 + 1.645(50)/\sqrt{a}$  or  $a = 67.65$ . From the equation for the prior mode,  $b = 67.65/50 = 1.353$ . The actual 95th percentile of a  $\text{Gamma}(67.65, 1.353)$  is 60.4, so the normal approximation is quite good.

**EXERCISE 5.34.** Carry out an analysis of the Ache data assuming the expert provides a “best guess” of  $\theta$  to be 1 armadillo and is 95% sure that the mean daily number of kills is under 1.5 armadillos. Give the estimated probability that a hunter comes home without bagging an armadillo.

**EXERCISE 5.35.** Suppose you require a prior for  $\theta$  with a mode of 2 and a 95th percentile of 4. Obtain an appropriate  $\text{Gamma}(a, b)$  prior and show that the normal approximation method is not very good. Then repeat with a mode of 40 and a 5th percentile of 25 and show that the normal approximation is reasonable.

### 5.3.3 Reference Priors

The Jeffreys' prior (cf. Section 4.7) for this problem is

$$p(\theta) \propto 1/\sqrt{\theta},$$

which is equivalent to an improper  $\text{Gamma}(0.5, 0)$  distribution. Just as for the conjugate prior, the posterior becomes

$$\theta | y \sim \text{Gamma}(n\bar{y} + 0.5, n).$$

The posterior mean is  $\bar{y} + 1/2n$ , the posterior standard deviation is  $\sqrt{\bar{y} + 1/2n}/\sqrt{n}$ . If the sample size is large, the large sample normal approximation kicks in and inferences will be nearly identical to those based on large sample MLE theory. This Jeffreys' prior can be approximated with a  $\text{Gamma}(0.5, 0.001)$  prior.

### 5.3.4 Two-Sample Poisson Data

We now compare independent Poissons with different rates (means).

**EXAMPLE 5.3.2.** The Nurses' Health Study (Colditz et al., 1990) sought to estimate and compare the rates of breast cancer for 50- to 59-year-old postmenopausal women who were current users of estrogen replacement therapy (group 1) and who were not (group 2). The data are given in Table 5.6. Do you find it disturbing that we would try to analyze such an important issue with only two observations: one on each group? What happened to random sampling?

Table 5.6: Breast cancer data.

Group	Cases	Person-years
1-Hormone Therapy	123	46,524
2-None	288	145,159

There is a funny thing about taking random samples from a Poisson distribution. You don't always have to do it to get good estimates. Let's think about BBs hitting our teacher's table again. Would it make any sense to cut the table up into 10 pieces and make 10 smaller tables out of it? Then we could get 10 observations on the Poisson, rather than one. No, what matters in terms of getting a good estimate of the rate at which BBs hit the table is the total area of table surface. It does not even matter if we make 10 tables of the same size, as long as their collective area is the same as the original table.

In Example 2.3.2 we mentioned that for  $y_1, \dots, y_n \stackrel{iid}{\sim} \text{Bern}(\theta)$  we get  $\sum y_i \sim \text{Bin}(n, \theta)$ . We also know that we can estimate  $\theta$  well from one binomial as long as the number of trials is large. Similarly, if we have  $n$  independent Poissons, say  $y_i \sim \text{Pois}(\theta_i)$ , then  $\sum y_i \sim \text{Pois}(\sum \theta_i)$ . But what we really are interested in with our BBs and tables is  $y_i \sim \text{Pois}(\theta M_i)$  where  $M_i$  is the size of the table. Then  $\sum y_i \sim \text{Pois}(\theta \sum M_i)$ . To analyze these data we only need to know  $\sum y_i$  and  $\sum M_i$ . In the breast cancer data,  $\sum M_i = 46,524$  for group 1 and  $\sum M_i = 145,159$  for group 2. Drawing an analogy with binomials, we have very large sample sizes but we *need* very large sample sizes because the event we are looking for is quite rare. A more precise analogy is that if BBs are only projected from outer space by alien youth, hitting our teacher's table will be rare. We will need a big table to catch an adequate number of BB strikes. But more importantly, if we want to compare BB strikes for Math and English classes, although it is important to have big tables so we get to see a fair number of strikes, it is even more important to the analysis to adjust for the size of the Math table relative to the English table. For breast cancer, group 2 has a "table" more than three times as large as group 1.

With these caveats, it is straightforward to generalize from one sample to two independent samples,

$$y_1 | \theta_1 \sim \text{Poisson}(\theta_1 M_1) \quad \perp \!\!\! \perp \quad y_2 | \theta_2 \sim \text{Poisson}(\theta_2 M_2).$$

We assume independent conjugate priors

$$\theta_1 \sim \text{Gamma}(a_1, b_1) \quad \perp \!\!\! \perp \quad \theta_2 \sim \text{Gamma}(a_2, b_2)$$

and, using a minor modification of the argument in Subsection 5.3.1, we get independent posteriors

$$\theta_1|y_1 \sim \text{Gamma}(a_1 + y_1, b_1 + M_1) \quad \perp\!\!\!\perp \quad \theta_2|y_2 \sim \text{Gamma}(a_2 + y_2, b_2 + M_2).$$

In addition to estimating the individual rates,  $\theta_1$  and  $\theta_2$ , effect measures commonly used to compare two Poisson populations are the difference of rates (means),  $\theta_1 - \theta_2$ , and the relative rates (means)  $\theta_1/\theta_2$ .

We should also mention that incorporating “table size” is going to change the process of prior elicitation. Probably the best way to approach it is to pick a standard size table, or cookie, or time period and elicit all prior information for that standard. Then the constants  $M_j$  should be defined as multiples of the standard unit and  $\theta_j$  will be the rate relative to the standard unit.

**EXAMPLE 5.3.2 CONTINUED.** Here  $y_j$ ,  $M_j$ , and  $\theta_j$  denote the number of new cases, the total person-years at risk *in thousands*, and the breast cancer incidence rate per 1,000 person-years in group  $j$ . Since we have no access to breast cancer experts, we use reference gamma priors on the incidence rates, i.e.,  $a_1 = a_2 = 0.5$ ,  $b_1 = b_2 = 0.001$ . Based on this cohort, the posteriors are

$$\theta_1|y_1 \sim \text{Gamma}(123.5, 46.525) \quad \perp\!\!\!\perp \quad \theta_2|y_2 \sim \text{Gamma}(288.5, 145.160),$$

so there was an estimated rate of 2.65 breast cancer cases per 1,000 person-years with hormone replacement therapy and 1.99 without. These are posterior means but they are virtually identical to the posterior medians. The 95% PIs are (2.21, 3.14) and (1.765, 2.225), respectively. The probability intervals have very little overlap.

The posterior median and 95% probability interval for the rate ratio comparing current hormone users to non-users are 1.33 and (1.08, 1.64). Moreover,  $\Pr[\theta_1 > \theta_2|y_1, y_2] > 0.99$ , so we are confident that the incidence of breast cancer was higher for 50- to 59-year-old postmenopausal women who were current users of estrogen replacement therapy. Although there is evidently an increased risk associated with estrogen replacement, the baseline risk of about 2 cancers per 1,000 person years is quite low as is the treatment group risk of 2.65. Even knowing the exact posterior distribution, these numbers are easier to obtain by simulation.

**EXERCISE 5.36.** Modify the WinBUGS code given below to reanalyze the breast cancer data using an informative prior where the modes for  $\theta_1$  and  $\theta_2$  are 3 and 1, respectively, and with 95 percentiles of 10 and 7, respectively. Are the results substantially different?

```
model{
  y1 ~ dpois(lambda1)
  y2 ~ dpois(lambda2)
  lambda1 <- theta1*M1
  lambda2 <- theta2*M2
  theta1 ~ dgamma(0.5,0.001)
  theta2 ~ dgamma(0.5,0.001)
  r1 <- 1000*theta1
  r2 <- 1000*theta2
  RateRatio <- theta1/theta2
  test <- step(RateRatio-1)
}
list(y1=123, M1=46524, y2=288, M2=145159)
list(theta1=1, theta2=1)
```

**EXERCISE 5.37.** Re-analyze the breast cancer data using the reference prior but now change the data so that  $M_1 = 93,000$  and  $M_2 = 290,000$ , that is, with about twice the amount of person time. Compare results with those obtained earlier.

EXERCISE 5.38. Suppose we have completely independent count data

$$y_{11}, \dots, y_{1n_1} | \theta_1 \stackrel{iid}{\sim} \text{Pois}(\theta_1) \quad \perp \quad y_{21}, \dots, y_{2n_2} | \theta_2 \stackrel{iid}{\sim} \text{Pois}(\theta_2)$$

with priors

$$\theta_1 \sim \text{Gamma}(a_1, b_1) \quad \perp \quad \theta_2 \sim \text{Gamma}(a_2, b_2).$$

Derive the analytical form of the posterior distribution of  $(\theta_1, \theta_2)$  and characterize it as a well-known distribution. Also identify the joint posterior when independent Jeffreys priors are used for  $\theta_1$  and  $\theta_2$ . Finally, write WinBUGS code to handle the model, including making inferences about the risk difference and risk ratio.

#### 5.4 Sample Size Determination\*

Sample size determination is a major component of experimental design. Enrolling too few experimental units may lead to an inconclusive study, while over-enrollment constitutes a drain on limited resources. We detail two simulation-based approaches to Bayesian sample size determination, one based on predictive probability and the other on average power. These approaches are useful in a wide variety of settings, including any of the models presented in this chapter and can also aid in the design of studies involving regression data. The methods are similar to those found in Geisser (1992).

Consider a situation where we want to compare two populations. We have data

$$y_{11}, \dots, y_{1n_1} | \theta_1 \stackrel{iid}{\sim} f_1(\cdot | \theta_1) \quad \perp \quad y_{21}, \dots, y_{2n_2} | \theta_2 \stackrel{iid}{\sim} f_2(\cdot | \theta_2)$$

with an effect measure  $g(\theta_1, \theta_2)$  that serves as the target parameter of interest. Our task is to find a sample size combination  $(n_1, n_2)$  that enables informed decisions to be made about  $g(\theta_1, \theta_2)$ .

Specifically, we might be interested in whether  $g(\theta_1, \theta_2) \equiv \theta_1 - \theta_2$  is close to zero or whether  $g(\theta_1, \theta_2) \equiv \theta_1 / \theta_2$  is close to one. Moreover, we are interested in collecting data that will tell us with high probability that, say,  $\theta_1 - \theta_2$  is close to zero when  $\theta_1 - \theta_2$  really is close to zero. More generally, the predictive approach involves finding an  $(n_1, n_2)$  pair that leads to a high probability in favor of

$$H : g(\theta_1, \theta_2) \in A,$$

when in fact  $H$  is true. The set  $A$  controls both the nature of the inference and the degree of inferential precision desired by the user. For example, with  $g(\theta_1, \theta_2) = \theta_1 - \theta_2$ , picking a small interval around zero determines that we want data to help correctly determine whether the  $\theta_i$ s are close but also how close they need to be. The “true” value of the effect measure is specified by picking true values  $\theta_1^T$  and  $\theta_2^T$  with  $g(\theta_1^T, \theta_2^T) \in A$ . Note that with  $g(\theta_1, \theta_2) \equiv \theta_1 - \theta_2$  and  $A$  a small interval around zero, we would probably pick  $\theta_1^T = \theta_2^T$  but, more than that, we need a specific value for this common number.  $\theta_1^T$  and  $\theta_2^T$  can be elicited from experts, based on historical data, or conservatively selected to ensure a successful study at a minimal effect size.

If  $H$  and the sampling model are true, the analysis of the future data should result in a high posterior probability of  $H$ , provided  $n_1$  and  $n_2$  are sufficiently large. That is, with future data  $y_i = (y_{i1}, \dots, y_{in_i})'$  sampled from the “true” model, and if the sample sizes are sufficiently large, we are likely to achieve

$$\Pr[g(\theta_1, \theta_2) \in A | y_1, y_2] \geq p_1$$

for some large  $p_1$  (such as 0.8, 0.9, or 0.95).

Since we haven’t seen the actual data, we regard  $\Pr[g(\theta_1, \theta_2) \in A | y_1, y_2]$  as random with a distribution that is induced by the “true” model for  $y_1$  and  $y_2$ . We want a sample size combination  $(n_1, n_2)$  that guarantees high predictive probability of achieving our goal, namely

$$\Pr\{\Pr[g(\theta_1, \theta_2) \in A | y_1, y_2] \geq p_1 | \theta_1^T, \theta_2^T\} \geq p_2$$

Table 5.7: Sample size adequacy for detecting  $RR > 1$  with two independent binomials having  $n = n_1 = n_2$ .

	$\theta_1^T = 0.25, \theta_2^T = 0.2$ $RR = 1.25$			$\theta_1^T = 0.3, \theta_2^T = 0.2$ $RR = 1.5$			$\theta_1^T = 0.35, \theta_2^T = 0.2$ $RR = 1.75$		
	$p_1$			$p_1$			$p_1$		
	0.80	0.90	0.95	0.80	0.90	0.95	0.80	0.90	0.95
$n = 50$	0.42	0.26	0.13	0.64	0.44	0.28	0.82	0.66	0.50
$n = 100$	0.49	0.32	0.18	0.80	0.66	0.48	0.92	0.85	0.75
$n = 150$	0.58	0.41	0.27	0.87	0.76	0.63	0.98	0.95	0.90
$n = 200$	0.62	0.45	0.32	0.93	0.86	0.76	0.99	0.99	0.96

for some  $p_2$ , say 0.8. So with the given sample sizes, we would be 80% sure that the data we collect would result in, say, a 95% posterior probability that  $H$  is true, when it is indeed true. The prior density  $p(\cdot)$  used for data analysis is often a reference prior when performing sample size calculations.

A Monte Carlo approximation is readily available. First simulate  $j = 1, \dots, s$  data sets with sample sizes  $(n_1, n_2)$  from the sampling model with  $\theta_1 = \theta_1^T$  and  $\theta_2 = \theta_2^T$ . For each data set  $j$ , sample from the posterior distribution to get  $(\theta_{1j}^k, \theta_{2j}^k)$ ,  $k = 1, \dots, m$ . The posterior probability of the effect measure being in  $A$  is numerically approximated as

$$\frac{1}{m} \sum_{k=1}^m I_A[g(\theta_{1j}^k, \theta_{2j}^k)], \quad (1)$$

for the  $j$ th data set. Overall, the sample size adequacy is then evaluated using

$$\frac{1}{s} \sum_{j=1}^s I_{(p_1, 1]} \left\{ \frac{1}{m} \sum_{k=1}^m I_A[g(\theta_{1j}^k, \theta_{2j}^k)] \right\}, \quad (2)$$

which is just the proportion of simulated data sets for which the posterior probability exceeded the criterion. If this number exceeds, say, 0.8 we are happy with the selected sample sizes. In general, the sample sizes are acceptable if the measure in (2) exceeds  $p_2$ . Naturally, we pick the smallest sample sizes that satisfy the  $p_2$  criterion.

**EXAMPLE 5.4.1.** In Example 3.1.3 and Subsection 5.1.3 we examined two independent binomial samples. Consider the task of finding a sample size combination that yields high predictive probability of a future study correctly concluding that a population risk ratio  $RR \equiv \theta_1/\theta_2 = g(\theta_1, \theta_2)$  exceeds 1. We calculated the predictive probability for uniform priors and equal sample sizes ( $n_1 = n_2 \equiv n$ ) of 50, 100, 150, and 200, with 3 values of  $p_1$  (0.80, 0.90, and 0.95), and with “true” risk ratios of 1.25, 1.50, and 1.75. For each sample size combination,  $s = 1000$  data sets were simulated with posterior approximations based on  $m = 2000$  iterates.

Table 5.7 gives values of (2). Lower predictive probabilities are attained for the smallest effect size considered (True  $RR = 1.25$ ). For instance, sample sizes of 200 in each group yield a predictive probability of only 0.62 when  $p_1 = 0.8$ , and the predictive probability drops to 0.32 when  $p_1 = 0.95$ . In contrast, larger effect sizes are likely to be detected with sample sizes of fewer than 200 in each group. Taking  $p_1 = p_2 = 0.8$ , a sample of between 100 and 150 subjects from each group is needed if the true RR is 1.5, while fewer than 100 total subjects need to be sampled if the true RR is 1.75.

**EXERCISE 5.39.** Write out all of the steps of a computational algorithm for calculating the predictive probability in the two-sample binomial setting when the risk ratio is the target effect measure,  $\theta_1^T = 0.2$ ,  $\theta_2^T = 0.4$ ,  $A = (1, \infty)$ , and  $p_1 = 0.8$ .

An alternative sample size procedure incorporates two separate prior distributions: the prior  $p(\theta_1, \theta_2)$  and a second “prior”  $p_*(\theta_1, \theta_2)$  that is concentrated on pairs  $(\theta_1, \theta_2)$  with  $g(\theta_1, \theta_2) \in A$ . Ideally,  $g(\theta_1, \theta_2) \in A$  would define the support of  $p_*$  but in practice  $p_*$  is often defined so that  $g(\theta_1, \theta_2) \in A$  merely has a high probability. These lead to alternative posteriors  $p(\theta_1, \theta_2 | y_1, y_2)$  and  $p_*(\theta_1, \theta_2 | y_1, y_2)$  and alternative marginal distributions for the data  $f(y_1, y_2)$  and  $f_*(y_1, y_2)$ .

With

$$\Pr[g(\theta_1, \theta_2) \in A | y_1, y_2] = \int I_A[g(\theta_1, \theta_2)] p(\theta_1, \theta_2 | y_1, y_2) d\theta_1 d\theta_2,$$

we define the *average power* as

$$\begin{aligned}\text{Average Power} &= E_* \{\Pr[g(\theta_1, \theta_2) \in A | y_1, y_2]\} \\ &= \int \Pr[g(\theta_1, \theta_2) \in A | y_1, y_2] f_*(y_1, y_2) dy_1 dy_2.\end{aligned}$$

The idea is that a sample size combination should be selected based on averaging over the parameters in  $H$  rather than selecting a parameter pair  $(\theta_1^T, \theta_2^T)$  in  $H$ . A sample size selection is acceptable if the average power exceeds some (high) pre-specified value.

The computational algorithm for calculating average power is similar to that for calculating a predictive probability. One key difference is that simulated data sets are generated conditional on values sampled from the prior  $p_*$ . Independent distributions for  $\theta_1$  and  $\theta_2$  used for simulating the data have been termed by Wang and Gelfand (2002) as *sampling priors*, to distinguish them from the prior  $p$  used for fitting the two-sample model to the simulated data. Sampling prior distributions are typically centered on presumed true values  $(\theta_1^T, \theta_2^T)$ , with some variability to account for the fact that the true values are uncertain.

A Monte Carlo integration proceeds by calculating (1) for each simulated data set  $j = 1, \dots, s$ , and then taking the average over the data sets as a numerical approximation to the average power achieved using  $(n_1, n_2)$ . This procedure is repeated using different sample size combinations to search for an  $(n_1, n_2)$  pair that yields sufficiently high power on average.

Other approaches to Bayesian sample size calculations are presented in the review paper by Adcock (1997) and references therein. Alternative criteria include average coverage, average length, and worst outcome.

**EXERCISE 5.40.** *Two-sample Binomial.* Calculate the predictive power of concluding that  $RR > 1$  in the two-sample binomial setting assuming a true risk ratio of 1.5 (with  $\theta_1^T = 0.3$  and  $\theta_2^T = 0.2$ ) for  $n_1 = n_2 = 50$ . Calculate the average power of concluding that  $RR > 1$  using as independent sampling priors a Beta distribution with mode 0.3 and 1st percentile 0.15 for  $\theta_1$ , and for  $\theta_2$ , a Beta distribution with mode 0.2 and 99th percentile 0.3. Use independent  $U(0, 1)$  priors for  $\theta_1$  and  $\theta_2$  for model fitting, and use 1,000 simulated data sets and sample 2,000 posterior iterates to generate numerical approximations. Repeat and compare for  $n_1 = n_2 = 200$ . Note that the computations can all be performed in R, but for practice you might use WinBUGS in tandem with R to fit each simulated data set. The code posted on our website that was used for Example 5.4.1 might help you get started.

---

## Chapter 6

---

# Simulations

---

Modern Bayesian analysis is typically performed by simulating the posterior distribution using Markov Chain Monte Carlo (MCMC) methods. Monte Carlo methods are a traditional name for simulation methods.

Historically, Bayesian methods were restricted by the need to perform integrations analytically. More recently, approximate Bayesian analysis has been performed by using numerical integrations (Naylor and Smith, 1982; Smith et al., 1985), by using the analytic Laplace approximation (Leonard, 1982; Tierney and Kadane, 1986; Kass et al., 1988), and by using Monte Carlo methods (Zellner and Rossi, 1984; Gelfand and Smith, 1990; Dellaportas and Smith, 1993). See Gelman et al. (2004, Chaps. 9–11) for a nice summary of these methods. We prefer Monte Carlo methods to Laplace approximations in regression problems because when performing many predictions, only a single Monte Carlo sample is necessary to perform all predictions, whereas the Laplace method requires a separate analytic approximation for each prediction. We prefer Monte Carlo methods to numerical integration because of their potential to deal with high-dimensional problems.

This chapter provides a short introduction to simulation, specifically, traditional simulation methods such as rejection sampling and importance sampling, as well as Markov chain theory and MCMC methods. Section 1 presents the basics of simulation and Section 2 presents the traditional methods of Acceptance-Rejection, and Importance Sampling. Section 3 presents a short version of the theory of Markov chains and its application to the methods of Gibbs Sampling, Metropolis Algorithm, and Slice Sampling, which are all used in WinBUGS. Section 3 also discusses adaptive rejection sampling and methods of assessing convergence. All readers should study Sections 6.3.0 and 6.3.5. Additional information on these topics can be found in Robert and Casella (2004), Chen, Shao, and Ibrahim (2000), Gilks, Richardson, and Spiegelhalter (1996), and Gamerman and Lopes (2006).

### 6.1 Generating Random Samples

Let  $Y$  be a random variable with strictly monotone cdf  $F(y)$  that has an inverse function  $F^{-1}(u)$ . By definition

$$\Pr(Y \leq y_0) = F(y_0).$$

Let  $U$  have a  $U(0, 1)$  distribution so that  $\Pr[U \leq u_0] = u_0$  and consider the random variable  $F^{-1}(U)$ .

$$\Pr[F^{-1}(U) \leq y_0] = \Pr[U \leq F(y_0)] = F(y_0),$$

so  $Y$  and  $F^{-1}(U)$  have the same distribution. In particular, if you have a way of generating random samples  $U_1, \dots, U_m$  from a  $U(0, 1)$  distribution, the independent random observations  $F^{-1}(U_1), \dots, F^{-1}(U_m)$  will all have the same distribution as  $Y$ .

**EXAMPLE 6.1.1. Exponentials, Gammas, and Betas.** If  $Y \sim \text{Exp}(\theta)$  then  $F(y) = 1 - e^{-\theta y}$  and  $F^{-1}(u) = -\log(1-u)/\theta$ . If we generate  $U \sim U(0, 1)$ ,  $-\log(U)/\theta \sim \text{Exp}(\theta)$ , since  $U \sim 1 - U$ . From inspecting the densities in Table 2.1, it is not difficult to see that  $\text{Exp}(\theta) = \text{Gamma}(1, \theta)$ .

A well-known property of Gamma distributions is that if  $Y_1, \dots, Y_n$  are independent  $\text{Gamma}(a_i, b)$ , then  $\sum_{i=1}^n Y_i \sim \text{Gamma}(\sum_{i=1}^n a_i, b)$ . Thus if  $U_1, \dots, U_n$  are iid  $U(0, 1)$ ,  $-\sum_{i=1}^n \log(U_i)/\theta \sim \text{Gamma}(n, \theta)$ .

Another well-known property of Gamma distributions is that if  $Y_1, Y_2$  are independent  $\text{Gamma}(a_i, b)$  distributions,  $Y_1/(Y_1 + Y_2) \sim \text{Beta}(a_1, a_2)$ . As such, it is a simple matter for us to simulate Beta distributions with integer parameters.

**EXERCISE 6.1.** Suppose  $y \sim \text{Beta}(a, 1)$ ,  $a > 0$ . Explain how to simulate  $y$  using its cdf.

**EXERCISE 6.2.** Suppose  $y$  is a random variable with cdf  $F(y) = 1 - e^{-\lambda y^\alpha}$  for  $y > 0$ ,  $\alpha > 0$ . We say  $y \sim \text{Weib}(\alpha, \lambda)$ . Explain how to simulate  $y$ . Note that  $y$  is a simple transformation of an exponential random variable. What is the transformation?

**EXAMPLE 6.1.2.** *Normals.* We now consider a convenient way to generate normal random variables. Take  $Z_1, Z_2$  iid  $N(0, 1)$ . Their distribution is rotationally symmetric about the origin in two-dimensional space. In other words, the density is constant on every circle. To specify the joint distribution of  $Z_1$  and  $Z_2$ , we need only specify how much density needs to be associated with every circle centered at the origin. In particular, all the points on a circle of radius  $\sqrt{Y}$  are all the  $Z_1, Z_2$  values satisfying  $Z_1^2 + Z_2^2 = Y$ . Knowing the distribution of  $Y$  and the rotational symmetry is enough to get us the entire distribution. By the definition of a chi-squared random variable,  $Z_1^2 + Z_2^2 \sim \chi^2(2) = \text{Gamma}(2/2, 1/2)$ , something we know how to simulate, cf. Example 6.1.1.

Take  $U_1 \sim U(0, 1) \perp\!\!\!\perp U_2 \sim U(0, 2\pi)$ . From Example 6.1.1

$$-2\log(U_1) \sim Z_1^2 + Z_2^2.$$

Intuitively, if  $U_2$ 's distribution is uniform on a circle of radius 1 about the origin, the corresponding points are  $(\cos(U_2), \sin(U_2))'$ , so the distribution of

$$[Z_1, Z_2]' \equiv [\sqrt{-2\log(U_1)} \cos(U_2), \sqrt{-2\log(U_1)} \sin(U_2)]'$$

is rotationally symmetric about the origin and puts the correct density on each circle. In other words,

$$Z_1, Z_2 \stackrel{iid}{\sim} N(0, 1). \quad (1)$$

This gives two  $N(0, 1)$  samples by sampling two uniforms. To sample  $Y_i \sim N(\mu, \sigma^2)$ , just take  $Y_i = \mu + \sigma Z_i$ .

**EXERCISE 6.3.** Using Proposition B.4, show that if  $r^2 \sim \chi^2(2) \perp\!\!\!\perp \phi \sim U(0, 2\pi)$ , then  $Z_1 = r\cos(\phi)$  and  $Z_2 = r\sin(\phi)$  are iid  $N(0, 1)$ . This result establishes (1). Remember that  $r^2 \equiv q$  is the original random variable and  $r = \sqrt{r^2} = \sqrt{q}$  is the transformation, so you need to find  $d\sqrt{r^2}/d(r^2) \equiv \sqrt{q}/dq$ .

**EXAMPLE 6.1.3.** *Multivariate Normals.* To sample a multivariate normal random vector  $y = (y_1, \dots, y_p)'$  with mean  $\mu = (\mu_1, \dots, \mu_p)'$  and covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix},$$

write  $\Sigma = AA'$ , generate  $z_1, \dots, z_p$  iid  $N(0, 1)$ , and compute

$$y = AZ + \mu$$

where  $Z = (z_1, \dots, z_p)'$ . One way to find an appropriate matrix  $A$  is to compute the spectral decomposition of  $\Sigma$ , that is,

$$\Sigma = PD(\lambda_i)P'$$

where  $D(\lambda_i)$  is a diagonal matrix with nonnegative diagonal elements  $\lambda_i$ . The  $\lambda_i$ s are eigenvalues of  $\Sigma$  and the columns of  $P$  are corresponding orthonormal eigenvectors. Many computer packages have the ability to compute these. Take  $A = PD(\sqrt{\lambda_i})$ .

**EXAMPLE 6.1.4.** *Multivariate Extension of Beta.* Beta distributions are useful for modeling prior beliefs about a single probability associated with whether an event occurs or does not occur. However, we often need to deal with multiple events. For example, we might wonder if a person's hair is blond, black, or brown. To examine this, we need probabilities for each of the outcomes. The Dirichlet distribution is a useful way of simultaneously specifying probabilities for multiple events.

Another well-known property of Gamma distributions is that if  $Y_1$ ,  $Y_2$ , and  $Y_3$  are independent  $\text{Gamma}(a_i, b)$  distributions,

$$\left( \frac{Y_1}{Y_1 + Y_2 + Y_3}, \frac{Y_2}{Y_1 + Y_2 + Y_3} \right) \sim \text{Dirichlet}(a_1, a_2, a_3).$$

For integer values of the  $a_i$ s, from Example 6.1.1 we know how to simulate the Gamma variables. Often we also use the notation

$$\left( \frac{Y_1}{Y_1 + Y_2 + Y_3}, \frac{Y_2}{Y_1 + Y_2 + Y_3}, \frac{Y_3}{Y_1 + Y_2 + Y_3} \right) \sim \text{Dirichlet}(a_1, a_2, a_3)$$

but in 3 dimensions this distribution has no density because the random variables are redundant. (They always add up to 1.) Also note that clearly

$$\frac{Y_1}{Y_1 + Y_2 + Y_3} \sim \text{Beta}(a_1, a_2 + a_3).$$

It is a simple matter to extend the Dirichlet distribution to allow for an arbitrary number of categories. If  $Y_1, Y_2, \dots, Y_k$  are independent  $\text{Gamma}(a_i, b)$  distributions, then define  $S = \sum_{i=1}^k Y_i$  and write

$$Z \equiv (Z_1, \dots, Z_k) \equiv \left( \frac{Y_1}{S}, \frac{Y_2}{S}, \dots, \frac{Y_k}{S} \right) \sim \text{Dirichlet}(a_1, \dots, a_k).$$

Since  $\sum_i Z_i = 1$ , there are really only  $k - 1$  free components of  $Z$ . From Example 6.1.1,

$$Z_i \sim \text{Beta}\left(a_i, \sum_{j \neq i} a_j\right).$$

**EXERCISE 6.4.** Use Proposition B.4 to transform  $Y_1, \dots, Y_k$  that are independent  $\text{Gamma}(a_i, b)$ s into the vector  $(Z_1, \dots, Z_{k-1}, S)'$ . Show that  $S$  is independent of the other variables by showing that the joint density for  $(Z_1, \dots, Z_{k-1}, S)$  is the product of a  $\text{Dirichlet}(a_1, \dots, a_k)$  density and a  $\text{Gamma}(\sum_i a_i, b)$  density. The Dirichlet density is an obvious extension of that given in Table 2.1.

**EXERCISE 6.5.** Place a prior on the vector of probabilities  $(\pi_1, \pi_2, \pi_3)$ . Here  $\pi_1$  is the prevalence of infection among individuals showing clinical symptoms while  $\pi_2$  is the prevalence among those who do not show clinical symptoms, so  $\pi \equiv \pi_1 + \pi_2$  is the prevalence of infection in a population of interest. For example, if an individual has the human immunodeficiency virus (HIV), they may or may not have acquired immunodeficiency syndrome (AIDS). (a) Construct a Dirichlet prior that has mean vector  $(2/10, 2/10, 6/10)$  and that has 95th percentile of 0.3 for  $\pi_1$ . Find  $a_1$  and  $a_2 + a_3$  by trial and error using BetaBuster. (b) What is the marginal prior for  $\pi_2$  and for  $\pi_3$ ? (c) Is it possible to find a Dirichlet prior with the given means and percentile that also has a 95th percentile for  $\pi_2$  of 0.5? Explain.

## 6.2 Traditional Monte Carlo Methods

In this section, we examine two methods that traditionally have been used for Monte Carlo computations: Acceptance-rejection sampling and importance sampling. In applications, these methods have largely been supplanted by the Markov chain Monte Carlo methods discussed in the next section. Their primary importance now is as a supplement to MCMC methods when individual distributions in a larger Markov chain are difficult to sample. Smith and Gelfand (1992) presented an introduction to the use of importance sampling and the rejection method for Bayesian analysis.

### 6.2.1 Acceptance-Rejection Sampling

We want to sample from a univariate distribution with density  $p(\theta)$  that is not recognizable as any easily sampled distribution. In fact, we may only know the kernel of the density, say  $p_*(\theta)$ , with  $p_*(\theta) \propto p(\theta)$ . (The posterior is often unrecognizable but it is proportional to the prior times the likelihood, two things we know.) To sample from  $p(\theta)$ , acceptance-rejection sampling uses another density that is “easy” to sample, say,  $q(\theta)$ , for which we can find a constant  $M$  that has  $p_*(\theta) \leq Mq(\theta)$  for all  $\theta$ . The function  $Mq(\theta)$  is called the upper envelope of the kernel. Ideally, it is as close as possible to  $p_*(\theta)$ . We return to the construction of  $q(\theta)$  and the determination of  $M$  shortly, but first assume that they are already known.

To sample from  $p(\cdot)$ , start by sampling  $\theta \sim q(\cdot)$  and independently  $U \sim U[0, 1]$ . The idea is that either  $\theta$  is acceptable as a sample from  $p(\cdot)$  or it is rejected and we start over.  $U$  determines whether  $\theta$  is acceptable. For given  $\theta$ ,  $Mq(\theta)U \sim U[0, Mq(\theta)]$ , so it is uniformly distributed below the envelope. We accept the sampled  $\theta$  if  $Mq(\theta)U < p_*(\theta)$  and reject it otherwise. Thus values of  $Mq(\theta)U$  between the upper envelope and the kernel are rejected. The probability of rejection is small if the area between the envelope and  $p_*(\cdot)$  is small.

Before showing that this method actually provides samples from  $p(\theta)$ , consider one particular method for constructing  $Mq(\theta)$ . Suppose that  $\ell(\theta) \equiv \log[p_*(\theta)]$  is concave. To pick  $Mq(\theta)$  find the mode of  $\ell(\theta)$ , pick points on either side of the mode, say,  $\tilde{\theta}_1$  and  $\tilde{\theta}_2$ , and envelop  $\ell(\theta)$  using the tangent lines at  $\tilde{\theta}_1$  and  $\tilde{\theta}_2$ . The tangent lines are

$$\gamma_i(\theta) \equiv \ell(\tilde{\theta}_i) + \dot{\ell}(\tilde{\theta}_i)(\theta - \tilde{\theta}_i)$$

where  $\dot{\ell}(\theta)$  is the derivative. Because  $\ell(\theta)$  is concave, for  $i = 1, 2$ ,  $\ell(\theta) \leq \gamma_i(\theta)$ . Exponentiating, we get

$$p_*(\theta) = e^{\ell(\theta)} \leq e^{\gamma_i(\theta)}.$$

This allows us to pick our envelope function as

$$Mq(\theta) = \min \left\{ e^{\gamma_1(\theta)}, e^{\gamma_2(\theta)} \right\}.$$

We have defined the product  $Mq(\theta)$  but for acceptance-rejection sampling, we need to actually know the density  $q(\theta)$ . As seen in Exercise 6.6, it is easy to find where the two tangent lines intersect, say at  $\theta_*$ , so

$$Mq(\theta) = \begin{cases} e^{\gamma_1(\theta)} & \text{if } \theta \leq \theta_* \\ e^{\gamma_2(\theta)} & \text{if } \theta \geq \theta_* \end{cases}.$$

Also as in Exercise 6.6, it is not difficult to integrate under this curve to find  $M$  and thus  $q(\theta)$ . Finally, Exercise 6.6 establishes that it is easy to sample from  $q(\theta)$ .

**EXERCISE 6.6.** Let  $\gamma_i(\theta) = a_i + b_i\theta$ . We know by construction that  $b_1 > 0$  and  $b_2 < 0$ . (a) Find  $\theta_*$  by setting  $\gamma_1(\theta_*) = \gamma_2(\theta_*)$  and solving. (b) Integrate  $Mq(\theta)$  over  $(-\infty, \infty)$  to determine  $M$  as a function of, say,  $\theta_*$  and the  $a_i$ s and  $b_i$ s. (c) Obtain the cdf based on the density  $q(\cdot)$ , say  $Q(v) \equiv \int_{-\infty}^v q(\theta)d\theta$ . Do this first for  $v \leq \theta_*$ , and then for  $v > \theta_*$ . Calculate the latter as  $\int_{-\infty}^{\theta_*} q(\theta)d\theta +$

$\int_{\theta_*}^v q(\theta) d\theta$ . (d) Finally, solve  $Q(v) = u$  for  $v$  so that  $v = Q^{-1}(u)$ . Thus if we sample  $U \sim U[0, 1]$ , we have  $Q^{-1}(U) \sim q(\cdot)$ . Carefully organize your presentation of the numerous steps involved.

**EXERCISE 6.7.** Suppose  $y|\theta \sim \text{Pois}(\theta)$  and  $\theta \sim \text{Gamma}(10, 0.5)$ . If  $y = 15$  is observed, develop a method for sampling from the posterior. Find the explicit form of the envelope function  $Mq(\theta)$ , including the point of intersection of the two tangent lines,  $\theta^*$ , corresponding to the tangent lines at  $\tilde{\theta}_1 = 10$  and  $\tilde{\theta}_2 = 22$ . Explain how you would find  $M$  and give an explicit algorithm for the accept-reject algorithm applied to this problem. You need not explicitly find  $M$ .

Finally, we establish the validity of the acceptance-rejection procedure. Define  $K(\theta) = p_*(\theta)/Mq(\theta)$ . Let  $A$  denote the event that  $\theta$  is accepted. Also let  $c_*$  be the constant that satisfies  $p(\theta) = c_* p_*(\theta)$ . We show that  $\Pr(\theta \leq v | A)$  equals the cdf corresponding to the density  $p(\cdot)$ .

$$\begin{aligned}\Pr(\theta \leq v | A) &= \frac{\Pr(\theta \leq v \text{ and } A)}{\Pr(A)} \\ &= \frac{\Pr[\theta \leq v \text{ and } Mq(\theta)U \leq p_*(\theta)]}{\Pr[Mq(\theta)U \leq p_*(\theta)]} \\ &= \frac{\Pr[\theta \leq v \text{ and } U \leq p_*(\theta)/Mq(\theta)]}{\Pr[U \leq p_*(\theta)/Mq(\theta)]} \\ &= \frac{\Pr[\theta \leq v \text{ and } U \leq K(\theta)]}{\Pr[U \leq K(\theta)]} \\ &= \frac{\int_{-\infty}^v \left[ \int_0^{K(\theta)} 1 du \right] q(\theta) d\theta}{\int_{-\infty}^{\infty} \left[ \int_0^{K(\theta)} 1 du \right] q(\theta) d\theta} \\ &= \frac{\int_{-\infty}^v K(\theta) q(\theta) d\theta}{\int_{-\infty}^{\infty} K(\theta) q(\theta) d\theta} \\ &= \frac{\int_{-\infty}^v [p_*(\theta)/Mq(\theta)] q(\theta) d\theta}{\int_{-\infty}^{\infty} [p_*(\theta)/Mq(\theta)] q(\theta) d\theta} \\ &= \frac{\int_{-\infty}^v c_* p_*(\theta) d\theta}{\int_{-\infty}^{\infty} c_* p_*(\theta) d\theta} \\ &= \int_{-\infty}^v p(\theta) d\theta.\end{aligned}$$

The last equality holds because the integral of  $p(\cdot)$  in the denominator is 1.

### 6.2.2 Importance Sampling

A random sample  $\theta^1, \dots, \theta^s$  from the posterior distribution can be viewed as an approximation to the posterior that, for  $k = 1, \dots, s$ , takes the value  $\theta^k$  with probability  $1/s$ . Importance sampling also provides an approximation to the posterior distribution that is discrete but one with unequal probabilities. It takes on values  $\theta^k$  with probability  $w_k$ . More formally, importance sampling provides numerical approximations to posterior integrals of the form

$$\int h(\theta) p(\theta|y) d\theta = \frac{\int h(\theta) L(\theta|y) p(\theta) d\theta}{\int L(\theta|y) p(\theta) d\theta}. \quad (1)$$

for some function  $h(\cdot)$ .

In importance sampling, one chooses a *known* density function  $q(\theta)$  that is easy to sample. The procedure works best if  $q(\theta)$  is similar in shape to the known kernel of the posterior  $L(\theta|y)p(\theta)$

with tails that do not decay more rapidly than the tails of the posterior. Sample  $\theta^1, \dots, \theta^s$  from the distribution with density  $q(\theta)$ . Define

$$\tilde{w}(\theta) \equiv \frac{L(\theta|y)p(\theta)}{q(\theta)}$$

and for  $k = 1, \dots, s$ ,

$$w_k \equiv \tilde{w}(\theta^k) / \sum_{j=1}^s \tilde{w}(\theta^j).$$

By design,  $\sum_k w_k = 1$ . The discrete approximation to the posterior takes the value  $\theta^k$  with probability  $w_k$ .

To see that this discrete approximation to the posterior provides accurate estimates of integrals like (1), rewrite (1) as

$$\int h(\theta)p(\theta|y)d\theta = \frac{\int h(\theta)[L(\theta|y)p(\theta)/q(\theta)]q(\theta)d\theta}{\int [L(\theta|y)p(\theta)/q(\theta)]q(\theta)d\theta} = \frac{\int h(\theta)\tilde{w}(\theta)q(\theta)d\theta}{\int \tilde{w}(\theta)q(\theta)d\theta}.$$

Applying the Law of Large Numbers, for large  $s$

$$\sum_{j=1}^s h(\theta^j)\tilde{w}(\theta^j)/s \doteq \int h(\theta)\tilde{w}(\theta)q(\theta)d\theta$$

and

$$\sum_{j=1}^s \tilde{w}(\theta^j)/s \doteq \int \tilde{w}(\theta)q(\theta)d\theta,$$

so applying the discrete approximation to  $E[h(\theta)|y]$  gives

$$\hat{\theta}_h \equiv \sum_{j=1}^s h(\theta^j)w_j = \frac{\sum_{j=1}^s h(\theta^j)\tilde{w}(\theta^j)/s}{\sum_{j=1}^s \tilde{w}(\theta^j)/s} \doteq \frac{\int h(\theta)\tilde{w}(\theta)q(\theta)d\theta}{\int \tilde{w}(\theta)q(\theta)d\theta} = \int h(\theta)p(\theta|y)d\theta.$$

To avoid any difficulties with the unequal weights in this discrete approximation to the posterior, we can obtain an approximate random sample from the posterior by taking a new Monte Carlo sample from this discrete distribution. The process is called *Sampling Importance Resampling* or *SIR*. Thus if a SIR sample is represented as  $(\theta^{*1}, \dots, \theta^{*m})$ , we can use this sample to make full inferences about  $h(\theta)$  by calculating  $h(\theta^{*1}), \dots, h(\theta^{*m})$  and obtaining a smoothed histogram, the mean, standard deviation, and quantiles in the usual way. (SIR is also the acronym for *sliced inverse regression* and we have used it as an acronym for the *standard improper reference* prior for normal data.)

A standard importance distribution  $q(\cdot)$  is a multivariate normal or multivariate Student distribution (see Exercise 9.3) with mean (location) equal to the posterior mode or MLE of  $\theta$  and covariance matrix (dispersion) equal to a scaled version of the Fisher Observed Information matrix's inverse (see Section 4.10). The importance distribution is taken to have heavier tails than the actual posterior because otherwise the weights  $w_k$  may get large when  $\theta^k$  is observed in an extreme tail of  $q(\theta)$ . Suppose an unusual value  $\theta^k$  occurs with very small  $q(\theta^k)$ . If the tails of  $q$  are much lighter than the kernel of the posterior,  $\tilde{w}(\theta^k) = p(\theta^k)L(\theta^k|y)/q(\theta^k)$  will be large and the weight given to  $\theta^k$  will be large. This results in unstable approximations to (1). We could have, say,  $w_k = 0.6$ , which would obviously result in a very poor discrete approximation to a continuous posterior. In theory, these issues take care of themselves, but in practice, we want to eliminate the possibility that outliers in the simulation sample, i.e., unusual values  $\theta^k$ , are given high weight. See Christensen (1997, Sec. 13.4) for details on applying importance sampling to logistic regression. Smith and Gelfand (1992) used the prior distribution as an importance function. This is often easier to sample, but it is unlikely to mimic the posterior as well.

EXERCISE 6.8. Suppose that  $y_1, y_2, \dots, y_n$  are iid  $\text{Pois}(\theta)$  and assume a Jeffreys' prior. Explain in detail how you would implement an importance sampling algorithm for obtaining the posterior probability that  $\theta > 20$ .

EXERCISE 6.9. Suppose that  $y_1, y_2, \dots, y_n$  are iid  $\text{Gamma}(\alpha, \beta)$  and assume  $p(\alpha, \beta) = p(\alpha)p(\beta)$ , where  $\beta \sim \text{Gamma}(a, b)$  and  $\alpha \sim \text{Gamma}(1, 1)$ . The prior mean for  $\alpha$  is one, so we are in some sense centering the prior on the exponential distribution but allowing departures from it. Explain in detail how you would implement an importance sampling algorithm for obtaining full inferences about the mean of the  $y$ s,  $\alpha/\beta$ .

### 6.3 Markov Chain Monte Carlo

The idea of Markov Chain Monte Carlo is to define a sequence of random vectors  $\theta^1, \theta^2, \theta^3, \dots$  in which the distribution of  $\theta^k$  near the beginning of the sequence can be just about anything but in which the distributions eventually settle down to the posterior distribution. Thus, if  $\theta^k$  has a marginal density  $q_k(\theta)$ , as  $k$  gets large, these densities approach the *posterior density*  $p(\theta|y)$ , which for this section we will abbreviate as  $p(\theta)$ . Specifically, the sequence of  $\theta^k$ 's is a Markov chain as defined in Subsection 6.3.1. Markov chains have other useful applications but our interest is restricted to sampling from the joint posterior. Under mild conditions, Markov chain theory indicates that if  $k$  is large, the  $\theta^k$ 's are (approximately) identically distributed with density  $p(\theta)$ . However, the  $\theta^k$ 's are not typically independent, so the  $\theta^k$ 's do not constitute an approximate random sample from the posterior. Nonetheless, a version of the Law of Large Numbers, sometimes called an *Ergodic Theorem*, applies to these sequences. Under some conditions, if  $\theta^1, \dots, \theta^s$  are sampled from a Markov chain and  $h$  is a function with finite expectation under the posterior distribution, then with probability one

$$\lim_{s \rightarrow \infty} \frac{1}{s} \sum_{j=1}^s h(\theta^j) = \int h(\theta) p(\theta) d\theta.$$

Thus we can approximate probabilities and expected values relative to the posterior distribution just by taking the sample mean of appropriate functions of the  $\theta^k$ 's.

*Burning-in* can improve the approximations. Intuitively, observations obtained after the chain has settled down to the posterior will be more useful in estimating probabilities and expectations for  $p(\theta)$ . If we throw out the early observations, taken while the process was settling down, the remainder of the process should be a very close approximation to one in which every observation is sampled from the posterior. Dropping the early observations is referred to as using a *burn-in* period. With simple statistical models, we might run the chain for 6,000 or 11,000 observations with a burn in of 1,000 observations, thus using the last 5,000 or 10,000 samples to estimate probability integrals associated with the posterior distribution. More complicated probability models typically require longer chains and burn-ins. Subsection 6.3.5 discusses checking on whether the chain has settled down. Without getting technical, we refer to the process of settling down as achieving stationarity, cf. Christensen (2001a, Section 4.1).

Given an observed sequence, a plot of the pairs  $(k, \theta^k)$  is known as a history. Figure 6.1 shows histories of four pairs of chains with each pair started at distinct initial values. Figure 6.1(a) shows the behavior we like to see. After a burn-in of 5,000 iterations, each chain has settled down nicely. Figure 6.1(b) shows typical behavior during the burn-in period. The two chains differ markedly in the initial stages but by 750 iterations we see that they are settling down. Figure 6.1(c) shows chains that are nowhere near settling down between 501 and 1,500 iterations, and Figure 6.1(d) shows the same chains having settled down nicely between 50,000 and 100,000 iterations. Although they have certainly converged by 50,000 iterations, observe the “waviness” of the histories compared with those in Figure 6.1(a). This is due to autocorrelation, which is discussed later.

Usually, even after the burn-in phase, the iterates are correlated. They are eventually identically distributed but not independent. If  $s = 10,000$  and we have correlation 1 between all pairs through

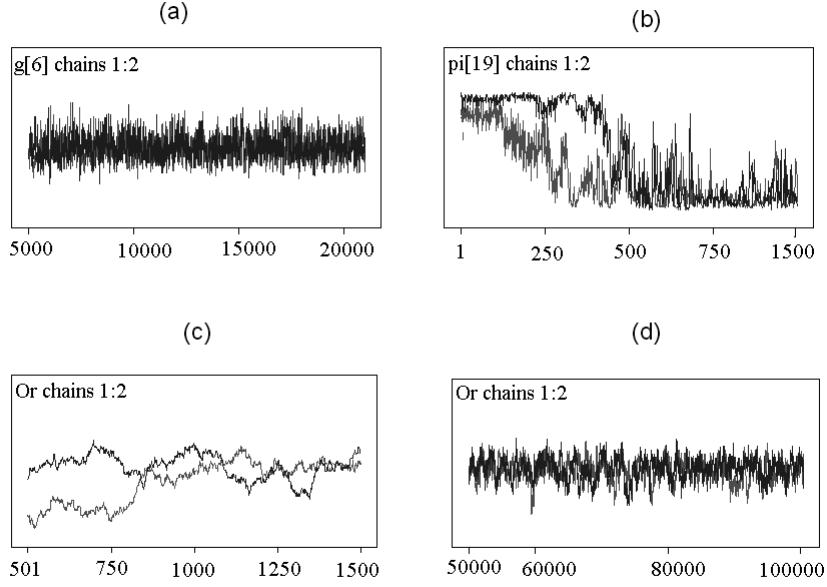


Figure 6.1: Histories of Markov chains with two starting values: (a) Chains that have converged after 5,000 iterations, (b) chains that diverge and then converge by 1,000 iterations, (c) chains not converging and exhibiting high autocorrelation, and (d) chains from (c) that ultimately converge sometime before 50,000 iterations, but that exhibit high autocorrelation.

time, then the effective Monte Carlo sample size is 1. This extreme case is unlikely ever to happen, but it illustrates a point. With iid sampling under typical circumstances, one can probably get reasonable approximations from a few thousand samples. If the samples are identically distributed but not independent, to attain sufficient accuracy in our numerical approximation to the posterior, we may need to take much larger sample sizes. Fortunately, computers are well equipped for that.

*Thinning* is a process used to make the observations more nearly independent, hence more nearly a random sample from the posterior distribution. Frankly, after a burn-in, there is not much point in thinning unless the correlations are extremely large. If there is a lot of correlation between adjacent observations, a larger overall MC sample size is needed to achieve reasonable numerical accuracy, in addition to needing a much longer burn-in. To check the level of dependence, look at the estimated *autocorrelation function (ACF)*. This is a function of the integers  $j$  that gives the estimated correlation between  $\theta^k$  and  $\theta^{k+j}$ . After a burn-in, this correlation should depend on the *lag*  $j$ , but not on  $k$ . It is computed as the sample correlation between the pairs  $(\theta^k, \theta^{k+j})$ ,  $k = 1, \dots, s - j$ . If the autocorrelations are near zero except for, say, the first two,  $j = 1, 2$ , then we could thin by taking every third  $\theta^k$ . That is, our sample could be  $\theta_{3k}$ ,  $k = 1, \dots, s$ , after the burn-in. This sample should be nearly uncorrelated but it throws away information. Unless there is severe autocorrelation, e.g., high correlation even with, say  $j = 30$ , we don't believe that thinning is worthwhile.

Figure 6.2 gives histories of two chains on the left and their corresponding autocorrelations on the right. The history and ACF on the top both look good. The ACF rapidly approaches zero and stays there. The autocorrelation function in Figure 6.2(d) dies out very slowly with autocorrelation of at least 0.5 even at a lag of 50 iterations. In time domain analysis of time series, such autocorrelation functions are taken as evidence of non-stationarity, cf. Christensen (2001a, Section 5.6). The waviness of the history plot indicates that iterations near one another are nearly the same. This waviness and the huge autocorrelations indicate the need for a larger Monte Carlo sample size.

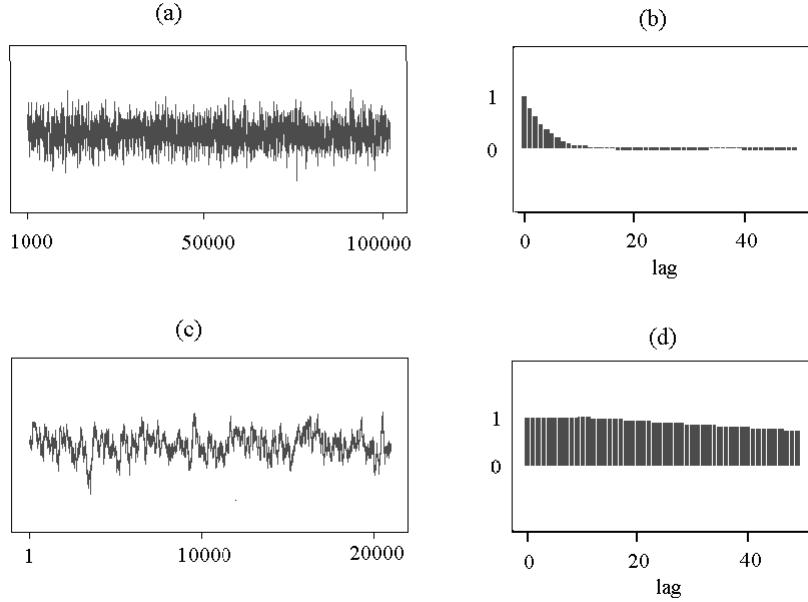


Figure 6.2: *Histories of Markov chains (left panels) and corresponding autocorrelations by lag (right panels): Panel (a) shows the history of a chain with strong autocorrelation as exhibited by the corresponding ACF in panel (b), and panel (c) shows the history of a chain with some autocorrelation as exhibited in panel (d).*

As discussed in the next section, the sequence (chain) should eventually settle down to be approximately stationary but these plots suggest that the convergence to stationarity is occurring very slowly or not at all. (The conditions that imply convergence may not hold.) We ran the chain from panel (c) out to 100,000 iterations and thinned by 20, and the ACF for the thinned chain looked very similar to the one in panel (b), and the history looked similar to the one in panel (a).

As a diagnostic to check on whether we have approximately identical distributions after burn-in, we might run several chains with  $\theta^1$  values chosen independently from an initial distribution. Such samples are independent and should give similar estimates for the posterior distribution. More on this is in Subsection 6.3.5.

The next subsection introduces some theory for Markov chains. The subsequent three subsections are devoted to specific methods for generating MCs that result in samples from the joint posterior after the burn-in phase, namely Gibbs sampling, the Metropolis algorithm, and Slice sampling. Casella and George (1992) presented an early introduction to the Gibbs sampler. When Gibbs sampling is applicable, it is almost always used. The Metropolis and Slice algorithms are more generally applicable, and are also used to augment the Gibbs sampler.

### 6.3.1 Markov Chains

Consider a sequence of random vectors  $\theta^1, \theta^2, \theta^3, \dots$ . This is a *Markov chain (MC)* if for any set  $A$ ,

$$\Pr(\theta^k \in A | \theta^1, \dots, \theta^{k-1}) = \Pr(\theta^k \in A | \theta^{k-1}). \quad (1)$$

In other words, what happens at step  $k$  depends only on what happened at step  $k - 1$ . Another way of thinking about it is that when you are at step  $k - 1$ , where you go depends only on where you are — it does not depend on how you got to where you are. This dependence of each new observation on only the previous observation is known as the *Markov property*.

Let  $q_1(\theta^1)$  be the initial density for  $\theta^1$ , let  $q_{k| \cdot}(\theta^k | \theta^1, \dots, \theta^{k-1})$  be the obvious conditional density, and as mentioned earlier,  $q_k(\theta^k)$  is the marginal density of  $\theta^k$ . From standard probability

$$\begin{aligned}\Pr(\theta^k \in A) &= \int_A q_k(\theta)d\theta \\ &= \int_A \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} q_{k| \cdot}(\theta^k | \theta^1, \dots, \theta^{k-1}) \cdots q_{2| \cdot}(\theta^2 | \theta^1) q_1(\theta^1) d\theta^1 d\theta^2 \cdots d\theta^k.\end{aligned}$$

If we have the Markov property (1), then we must have  $q_{j| \cdot}(\theta^j | \theta^1, \dots, \theta^{j-1}) = q_{j| j-1}(\theta^j | \theta^{j-1})$  for  $j = 2, 3, \dots$ , so a Markov chain has

$$\Pr(\theta^k \in A) = \int_A \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} q_{k| k-1}(\theta^k | \theta^{k-1}) \cdots q_{2| 1}(\theta^2 | \theta^1) q_1(\theta^1) d\theta^1 d\theta^2 \cdots d\theta^k.$$

Constructing a Markov chain is simple. All you have to do is specify the initial distribution  $q_1(\theta^1)$  and the conditional distributions  $q_{j| j-1}(\theta^j | \theta^{j-1})$  for  $j = 2, 3, \dots$ . To show that something is a Markov chain, all you have to do is show that the conditional distributions are of the form  $q_{j| j-1}(\theta^j | \theta^{j-1})$  for  $j = 2, 3, \dots$ . Moreover, it is simple to sample from an MC if you know  $q_1(\cdot)$  and all of the conditional densities. Generate  $\theta^1$  from  $q_1$ , then, since you know  $\theta^1$ , generate  $\theta^2$  from the appropriate conditional distribution, and continue.

A simplifying assumption that we make *henceforth* is that of *stationary transition probabilities*. This assumption states that the conditional distribution of going from step  $j - 1$  to step  $j$  is the same regardless of the value of  $j$ . If we write the densities more carefully as in Appendix B,  $q_{j| j-1}(\theta^j | \theta^{j-1}) \equiv q_{\theta^j | \theta^{j-1}}(u|v)$ . For this function not to depend on the steps  $j - 1$  and  $j$ , it must be some function  $q(u|v)$ , which we commonly write as  $q(\theta^j | \theta^{j-1})$  to help us keep track of where it is being applied. (Many discussions use the notation  $q(u|v) \equiv k(v,u)$ , which is called a *transition kernel*.) A Markov chain with stationary transition probabilities now has

$$\Pr(\theta^k \in A) = \int_A \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} q(\theta^k | \theta^{k-1}) \cdots q(\theta^2 | \theta^1) q_1(\theta^1) d\theta^1 d\theta^2 \cdots d\theta^k.$$

Although the transition probabilities do not depend on the step of the chain, the marginal distribution of a  $\theta^k$  typically depends on  $k$ , hence  $q_k(\theta^k)$  still indicates the marginal density at step  $k$ .

Historically, an interesting issue for Markov chains with stationary transition probabilities is studying the effect that the choice of an initial distribution  $q_1(\theta^1)$  has on the marginal distributions  $q_k(\theta^k)$ . It is the solution of this problem that makes Markov chains useful in statistical simulation.

One aspect of the solution involves the idea of a *stationary distribution*. A stationary distribution, say  $p(\cdot)$ , has the property for all  $k$  that

$$\Pr(\theta^k \in A) = \int_A p(\theta)d\theta. \tag{2}$$

In particular,  $q_k(\theta) = p(\theta)$  for all  $\theta$  and for all  $k$ , including  $k = 1$  which means we start the chain with a value taken from the stationary distribution itself. Under the assumption of a stationary distribution and using the law of total probability, we find

$$\begin{aligned}\int_A p(\theta)d\theta &= \Pr(\theta^k \in A) \\ &= \int \Pr(\theta^k \in A | \theta^{k-1}) p(\theta^{k-1}) d\theta^{k-1} \\ &= \int \left[ \int_A q(\theta^k | \theta^{k-1}) d\theta^k \right] p(\theta^{k-1}) d\theta^{k-1} \\ &= \int_A \left[ \int q(\theta^k | \theta^{k-1}) p(\theta^{k-1}) d\theta^{k-1} \right] d\theta^k.\end{aligned}$$

Since this holds for any set  $A$ , and since  $\theta^k$  and  $\theta^{k-1}$  are just dummy variables in the integrals, we must have

$$p(\theta) = \int q(\theta|\theta^{k-1})p(\theta^{k-1})d\theta^{k-1} = \int q(\theta|\theta^*)p(\theta^*)d\theta^*. \quad (3)$$

Thus a chain that satisfies (2) for all  $k$  must also satisfy (3).

The result that we really want is the converse. If we start our chain with  $q_1(\theta) = p(\theta)$ , and if our transition density satisfies (3), then (2) will be satisfied, namely, the chain will have stationary density  $p(\theta)$ . This is easily established by induction. Assume that the first iterate is taken from  $p(\theta)$  and that (3) holds. It follows from the law of total probability that

$$q_2(\theta^2) = \int q(\theta^2|\theta^1)p(\theta^1)d\theta^1 = p(\theta^2).$$

Then assuming that  $q_k(\theta) = p(\theta)$  as the induction hypothesis, exactly the same argument leads to the conclusion that  $q_{k+1}(\theta) = p(\theta)$  and we are done.

Henceforth we assume the existence of a stationary distribution. However, it is possible that there could be more than one density  $p$  that satisfies (3), hence more than one stationary distribution. Under relatively weak conditions, that cannot happen.

So far, we have not accomplished very much. If we construct a Markov chain that has the posterior  $p(\theta)$  as a stationary distribution, then if we begin the MC by sampling from  $p(\theta)$ , every subsequent observation will also be an observation with density  $p(\theta)$ . But if we knew how to sample from  $p(\theta)$ , we would not need any of this Markov chain machinery. The point is to run a Markov chain that starts arbitrarily but eventually gives us samples from the posterior. We need some powerful results to achieve this goal.

Under conditions discussed later, when a proper stationary distribution exists, it is unique, also, as  $k$  gets large, regardless of the initial distribution  $q_1(\theta^1)$ , the marginal distribution of the  $\theta^k$ 's settles down to the stationary distribution, and, finally, the Markov chain satisfies a version of the Law of Large Numbers. Rephrasing the middle result, as  $k$  gets large, the  $q_k(\theta^k)$  distribution approaches the  $p(\theta^k)$  distribution, that is, for large  $k$ ,

$$\Pr(\theta^k \in A) \doteq \int_A p(\theta)d\theta,$$

or more technically,

$$\lim_{k \rightarrow \infty} \Pr(\theta^k \in A) = \int_A p(\theta)d\theta. \quad (4)$$

In particular, by picking  $q_1$  to give probability one to the value  $\theta_*$ ,

$$\lim_{k \rightarrow \infty} \Pr(\theta^k \in A | \theta^1 = \theta_*) = \int_A p(\theta)d\theta. \quad (5)$$

This suggests that, rather than randomly picking where to start the chain, we can start it anywhere we want. Nonetheless, it seems obvious that the convergence should be faster if we can pick an initial distribution  $q_1$  that is somehow close to  $p$ , or an initial value  $\theta^1$  that is characteristic of  $p$ .

The key result is the Markov chain version of the Law of Large Numbers, sometimes called an *Ergodic Theorem*. It states that if  $\theta^1, \dots, \theta^s$  are sampled from the Markov chain and  $h$  is a function with finite expectation under the stationary distribution, then with probability one

$$\lim_{s \rightarrow \infty} \sum_{j=1}^s h(\theta^j)/s = \int h(\theta)p(\theta)d\theta. \quad (6)$$

Thus we can approximate probabilities and expected values relative to the stationary (in practice, posterior) distribution. This makes sense because for large  $k$ , the  $\theta^k$ 's are approximately identically distributed with density  $p(\theta)$ , although they are typically not independent.

The usefulness of all of this is that we can construct Markov chains with particular stationary transition probabilities that have the posterior distribution as their stationary distribution. Then, regardless of the initial distribution we choose, if we sample from the MC long enough, the samples will, to a good approximation, be identically distributed from the posterior and sample means converge with probability one to their corresponding posterior expectations. To construct such an MC, we need to establish that it has stationary transition probabilities, that the posterior satisfies equation (3), and a little more.

We now make some definitions and present formal conditions for (4), (5), and (6) to hold. We take most of our results from Tierney (1994). An MC with stationary distribution  $p(\theta)$  is *p-irreducible* if, for any initial value, there is positive probability of eventually reaching any set  $A$  for which  $\int_A p(\theta)d\theta > 0$ . A chain is *periodic* if it can only return to an initial set at regularly spaced times, e.g., on the 2nd, 4th, 6th, etc. iterations. Otherwise, it is *aperiodic*. A sufficient condition for aperiodicity is that  $\int_A q(\theta|\theta^1)d\theta > 0$  for all  $\theta^1$ , provided  $\int_A p(\theta)d\theta > 0$ , which means that it is possible to get to any set of interest in one transition, regardless of where the chain was started. This condition is also sufficient for *p*-irreducibility.

If a chain has a proper stationary distribution  $p$ , is *p*-irreducible, and is aperiodic, then not only is the stationary distribution unique, that is, no other choice for  $p$  can satisfy (3) for the given transition distribution  $q(\cdot|\cdot)$ , but (5) is satisfied.

Convergence theory for MCs involves assessing whether a chain will repeatedly return to any specified set with positive probability under  $p(\theta)$ . This is called *recurrence*. The form of recurrence we need is *Harris recurrence*. A chain is Harris recurrent if, for every starting value  $\theta^1 = \theta_*$  and any set  $A$  with positive probability under  $p(\theta)$ , the probability that  $A$  is revisited by the chain infinitely often is one. This guarantees, in theory if not in practice, *good mixing* of the chain. We want the MC to explore the entire support of the posterior (stationary) distribution, so it is important that every region of the support that has positive probability be visited infinitely often by the chain. A sufficient condition for Harris recurrence in an MC with stationary  $p$  is that it be *p*-irreducible and that  $\int_A p(\theta)d\theta = 0$  implies  $\int_A q(\theta|\theta^1)d\theta = 0$ , for all initial values  $\theta^1$ . This latter condition is referred to as the transition distribution being absolutely continuous with respect to the stationary distribution. The absolute continuity condition reduces to checking whether sets with posterior probability zero also have transition probability zero, regardless of the initial value.

A Markov chain is called ergodic if it is Harris recurrent and aperiodic. If we have an ergodic chain with stationary distribution  $p(\theta)$ , (4) holds, and if  $h(\cdot)$  is integrable with respect to  $p(\theta)$ , then (6) holds.

In practice, we check that  $\int_A p(\theta)d\theta = 0$  if and only if  $\int_A q(\theta|\theta^1)d\theta = 0$  for all  $\theta^1$ . If this is true, the MC is aperiodic, *p*-irreducible, and Harris recurrent, so all of (4), (5), and (6) hold.

MCs generated by Gibbs samplers, the Metropolis algorithm, and slice sampling all have the posterior satisfying (3) and are usually ergodic. We establish (3) for Gibbs sampling and the Metropolis algorithm in the next two subsections. Slice sampling is a special case of the Gibbs sampler. Details for establishing ergodicity can be found in Tierney (1994) and Robert and Casella (2004).

### 6.3.2 Gibbs Sampling

Gibbs sampling is a method for constructing a Markov chain that is extremely useful when one can isolate the conditional distribution of each parameter given all of the other parameters. The process involves obtaining samples from each conditional distribution in turn. More generally, it can be applied to sets of parameters. Temporarily treating vectors as row vectors, we illustrate the ideas for three blocks or subvectors, that is,  $\theta^k = (\theta_1^k, \theta_2^k, \theta_3^k)$ . The dimensions of each block are arbitrary. We construct the chain so that the posterior  $p(\theta) = p(\theta_1, \theta_2, \theta_3)$  is the stationary distribution. Here we have again dropped the explicit dependence on the data for brevity. Gibbs sampling is based on

sampling from the *full conditional distributions* determined by the posterior, i.e.,

$$p_{1|23}(\theta_1 | \theta_2, \theta_3), \quad p_{2|13}(\theta_2 | \theta_1, \theta_3), \quad p_{3|12}(\theta_3 | \theta_1, \theta_2).$$

To define the MC, first sample  $\theta^1$  from the initial distribution  $q(\theta_1, \theta_2, \theta_3) \equiv q_1(\theta)$ . This can be a one-point distribution, i.e., just pick a starting value, or, if making a random selection, it may be convenient to have the three blocks independent, thus sampling  $\theta^1 = (\theta_1^1, \theta_2^1, \theta_3^1)$  as

$$\theta_1^1 \sim q_1(\theta_1), \quad \theta_2^1 \sim q_2(\theta_2), \quad \theta_3^1 \sim q_3(\theta_3)$$

where we have implicitly redefined the  $q_1$  notation. The key to Gibbs sampling is that the transition probabilities are defined in terms of the full conditional distributions. The second complete step of the chain defines  $\theta^2$  in three phases. First,

$$\theta_1^2 | \theta_2^1, \theta_3^1 \sim p_{1|23}(\theta_1 | \theta_2^1, \theta_3^1),$$

then

$$\theta_2^2 | \theta_1^2, \theta_3^1 \sim p_{2|13}(\theta_2 | \theta_1^2, \theta_3^1),$$

and finally

$$\theta_3^2 | \theta_1^2, \theta_2^2 \sim p_{3|12}(\theta_3 | \theta_1^2, \theta_2^2).$$

In general, we sample  $\theta^k = (\theta_1^k, \theta_2^k, \theta_3^k)$  as

$$\theta_1^k | \theta_2^{k-1}, \theta_3^{k-1} \sim p_{1|23}(\theta_1 | \theta_2^{k-1}, \theta_3^{k-1}),$$

$$\theta_2^k | \theta_1^k, \theta_3^{k-1} \sim p_{2|13}(\theta_2 | \theta_1^k, \theta_3^{k-1}),$$

and

$$\theta_3^k | \theta_1^k, \theta_2^k \sim p_{3|12}(\theta_3 | \theta_1^k, \theta_2^k).$$

By construction, this defines a valid conditional distribution for transitioning from  $\theta^{k-1}$  to  $\theta^k$  that does not depend on  $k$ . In particular, the stationary transition distribution is

$$\begin{aligned} q(\theta^k | \theta^{k-1}) &\equiv q(\theta_1^k, \theta_2^k, \theta_3^k | \theta_1^{k-1}, \theta_2^{k-1}, \theta_3^{k-1}) \\ &\equiv p_{1|23}(\theta_1^k | \theta_2^{k-1}, \theta_3^{k-1}) p_{2|13}(\theta_2^k | \theta_1^k, \theta_3^{k-1}) p_{3|12}(\theta_3^k | \theta_1^k, \theta_2^k). \end{aligned} \tag{7}$$

With these transition distributions, the posterior is the stationary distribution, a fact that we will later illustrate for a two-block Gibbs sampler.

**EXAMPLE 6.3.1.** *Normal data with Independence Prior.* Suppose we have observations

$$y_1, \dots, y_n | \mu, \tau \stackrel{iid}{\sim} N(\mu, 1/\tau)$$

with prior

$$\mu \sim N(a, 1/b) \perp\!\!\!\perp \tau \sim \text{Gamma}(c, d).$$

As discussed in Subsection 5.2.4, the posterior density  $p(\mu, \tau | y)$  is not recognizable as any parametric form but we saw that

$$\mu | \tau, y \sim N[\hat{\mu}(\tau), 1/(n\tau + b)]$$

and

$$\tau | \mu, y \sim \text{Gamma}\left(c + \frac{n}{2}, d + \frac{1}{2} [n(\bar{y} - \mu)^2 + (n-1)s^2]\right).$$

To sample these distributions iteratively, we start with initial values  $(\mu^1, \tau^1)$ . To get off to a good start, these are often picked to be the modes of informative prior distributions. We then sample a new value  $\mu^2$  from the  $N[\hat{\mu}(\tau^1), 1/(n\tau^1 + b)]$  distribution followed by a sample  $\tau^2$  taken from the  $\text{Gamma}(c + n/2, d + [n(\bar{y}) - \mu^2]^2 + (n-1)s^2]/2)$  distribution. We continue in this fashion to obtain  $s$  iterates,  $\{(\mu^k, \tau^k) : k = 1, \dots, s\}$ . Moreover, as discussed in Chapter 3, for any function  $\gamma = g(\mu, \tau)$ , we can approximate the posterior  $p(\gamma|y)$  numerically by using the sample  $\{\gamma^k \equiv g(\mu^k, \tau^k) : k = 1, \dots, s\}$ .

The Gibbs sampler presupposes that we actually know how to sample from the full conditional distributions. Sometimes these distributions are recognizable, such as normal, beta, or gamma distributions, in which case sampling from them is easy. In any case, we can find the kernels of the full conditionals which lets us sample from the distribution using something like an *adaptive-rejection method*, a Metropolis method, or a Slice sampler. Adaptive-rejection sampling is a modification of the acceptance-rejection method of Section 2 but is more efficient as applied to Gibbs sampling and is implemented in WinBUGS when the full conditional is log concave. The other methods are discussed in the next two subsections and are also built into WinBUGS to handle full conditionals that are not log concave.

Adaptive-rejection sampling was developed by Gilks and Wild (1992). A simpler algorithm was used by Dellaportas and Smith (1993) for Bayesian generalized linear models, which generally have log concave posteriors provided the priors are log concave. They combined the Gibbs sampler with the envelope rejection method for sampling unrecognizable full conditionals. To sample an observation, say  $\theta^k$ , using the envelope method requires choosing two values, say  $\tilde{\theta}_j$ ,  $j = 1, 2$ , one on either side of the mode. Taking the  $\tilde{\theta}_j$ s to be the most recent values of  $\theta^{k-1}, \dots, \theta^1$  to fall on either side of the mode seems to work well.

**EXAMPLE 6.3.2. Weibull Data.** Feigl and Zelen (1965) present data on the survival times, measured in weeks, of patients who were diagnosed with leukemia. The patients were classified according to one of two characteristics of white blood cells. We only examine those whose blood is AG+. The sample consists of  $n = 17$  times in weeks from diagnosis to death: 65, 156, 100, 134, 16, 108, 121, 4, 39, 143, 56, 26, 22, 1, 1, 5, 65. Survival times are dealt with in detail in Chapters 12 and 13.

The data must be strictly positive, so we need models that are concentrated on  $(0, \infty)$ . One such model is the Weibull distribution. We denote

$$y_i \sim \text{Weib}(\alpha, \lambda)$$

if the density and cdf are

$$f_*(y_i | \alpha, \lambda) = \lambda \alpha y_i^{\alpha-1} e^{-\lambda y_i^\alpha}; \quad F_*(y_i | \alpha, \lambda) = 1 - e^{-\lambda y_i^\alpha}, \quad y_i > 0.$$

With data

$$y_1, \dots, y_n | \alpha, \lambda \stackrel{iid}{\sim} \text{Weib}(\alpha, \lambda),$$

the likelihood is

$$L(\alpha, \lambda) \propto \prod_{i=1}^n \lambda \alpha y_i^{\alpha-1} e^{-\lambda y_i^\alpha} = \lambda^n \alpha^n \left( \prod_{i=1}^n y_i \right)^{\alpha-1} \exp \left( -\lambda \sum_{i=1}^n y_i^\alpha \right).$$

Discussion of placing an informative prior on  $(\alpha, \lambda)$  is deferred to Chapters 12 and 13 but we select a prior with  $\lambda \sim \text{Gamma}(a, b)$ ,  $\alpha$  independent of  $\lambda$ , and  $\alpha$  with a prior density  $p_0(\alpha)$  having support  $(0, \infty)$ . The joint prior density has the form

$$p(\alpha, \lambda) \propto p_0(\alpha) \lambda^{a-1} e^{-\lambda b}.$$

The posterior density has the form

$$\begin{aligned} p(\alpha, \lambda | y) &\propto \lambda^n \alpha^n \left( \prod_{i=1}^n y_i \right)^{\alpha-1} \exp \left( -\lambda \sum_{i=1}^n y_i^\alpha \right) p_0(\alpha) \lambda^{a-1} e^{-\lambda b} \\ &\propto \lambda^{a+n-1} \exp \left[ -\lambda \left( b + \sum_{i=1}^n y_i^\alpha \right) \right] \alpha^n \left( \prod_{i=1}^n y_i \right)^\alpha p_0(\alpha). \end{aligned}$$

We know of no choice for  $p_0(\alpha)$  that makes the joint posterior recognizable.

Gibbs sampling for this problem involves an additional wrinkle. The conditional density for  $\lambda | \alpha, y$  is easily seen to be

$$p(\lambda | \alpha, y) \propto \lambda^{a+n-1} \exp \left[ -\lambda \left( b + \sum_{i=1}^n y_i^\alpha \right) \right],$$

so

$$\lambda | \alpha, y \sim \text{Gamma} \left( a+n, b + \sum_{i=1}^n y_i^\alpha \right).$$

However, the conditional density for  $\alpha | \lambda, y$  is

$$p(\alpha | \lambda, y) \propto \alpha^n \left( \prod_{i=1}^n y_i \right)^\alpha \exp \left[ -\lambda \sum_{i=1}^n y_i^\alpha \right] p_0(\alpha).$$

Again, we know of no choice for  $p_0(\alpha)$  that makes the conditional recognizable. However, we have discussed methods of sampling from an unknown distribution, so we would simply use one of them. The full conditional is log concave provided  $p_0(\alpha)$  is log concave, and so it is possible to sample from it using adaptive-rejection sampling. In fact, that is precisely what WinBUGS does for this problem.

Gibbs sampling again begins with an initial value for  $(\alpha, \lambda)$ , say  $(\alpha^1, \lambda^1)$ . It is easy to sample a new  $\lambda$  for given  $\alpha^1$  by sampling a  $\text{Gamma}(a+n, b + \sum_{i=1}^n y_i^{\alpha^1})$  variate. Call it  $\lambda^2$ . Next, sample a new value  $\alpha^2$  from the conditional distribution  $\alpha | \lambda^2, y$ , and continue until a sufficient number of samples have been taken. We thus obtain  $\{(\alpha^k, \lambda^k) : k = 1, \dots, s\}$ , which after a burn-in phase constitutes a sample from the joint posterior distribution.

**EXERCISE 6.10.** The following code provides an analysis for the data of Example 6.3.2 using independent gamma priors on the parameters and allows inferences about the median time to death and the 24-week survival rate, i.e.,  $1 - F_*(24 | \alpha, \lambda)$  where  $F_*$  is the cdf of the Weibull.

```
model{
  for(i in 1:n){ y[i] ~ dweib(alpha,lambda) }
  lambda ~ dgamma(1.53,26.3)
  alpha ~ dgamma(1,1)
  median <- log(2)/lambda
  S24 <- exp(-lambda*pow(24,alpha))
}
list(n=17,
      y=c(65,156,100,134,16,108,121,4,39,143,56,26,22,1,1,5,65))
list(lambda=0.05, alpha =1)
```

Run the code and make inferences about the median and 24-week survival times. Also check to see if the data might suggest that  $\alpha$  is near 1, which would imply that an exponential distribution might suffice as a model for the data.

EXERCISE 6.11. Show that the full conditional for  $\alpha$  in Example 6.3.2 is log concave provided  $p_0(\alpha)$  is log concave.

EXERCISE 6.12. Argue that the chain generated in Example 6.3.2 by sampling the full conditional for  $\lambda$  from the appropriate Gamma distribution, and the full conditional for  $\alpha$  using acceptance-rejection sampling, as described in Subsection 6.2.1, will result in samples from the joint posterior.

### 6.3.2.1 Proof that $p(\theta)$ is the Stationary Distribution in the Two-Block Case\*

We need to show that (3) holds for the two block case. In the two block case, using the same argument as in (7), the stationary transition density for Gibbs sampling is

$$\begin{aligned} q(\theta^k | \theta^{k-1}) &\equiv q(\theta_1^k, \theta_2^k | \theta_1^{k-1}, \theta_2^{k-1}) \\ &= p_{1|2}(\theta_1^k | \theta_2^{k-1}) p_{2|1}(\theta_2^k | \theta_1^k). \end{aligned}$$

In order to establish (3), we must show that

$$\int p(\theta^{k-1}) q(\theta^k | \theta^{k-1}) d\theta^{k-1} = p(\theta^k).$$

From the definition of the Gibbs sampler and using its transition probability density, the left hand side above is

$$\begin{aligned} &\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(\theta_1^{k-1}, \theta_2^{k-1}) p_{1|2}(\theta_1^k | \theta_2^{k-1}) p_{2|1}(\theta_2^k | \theta_1^k) d\theta_1^{k-1} d\theta_2^{k-1} \\ &= \int_{-\infty}^{\infty} \left[ \int_{-\infty}^{\infty} p(\theta_1^{k-1}, \theta_2^{k-1}) d\theta_1^{k-1} \right] p_{1|2}(\theta_1^k | \theta_2^{k-1}) p_{2|1}(\theta_2^k | \theta_1^k) d\theta_2^{k-1} \\ &= \int_{-\infty}^{\infty} p_2(\theta_2^{k-1}) p_{1|2}(\theta_1^k | \theta_2^{k-1}) p_{2|1}(\theta_2^k | \theta_1^k) d\theta_2^{k-1} \\ &= \int_{-\infty}^{\infty} p(\theta_1^k, \theta_2^{k-1}) p_{2|1}(\theta_2^k | \theta_1^k) d\theta_2^{k-1} \\ &= p_1(\theta_1^k) p_{2|1}(\theta_2^k | \theta_1^k) \\ &= p(\theta^k), \end{aligned}$$

as we intended to show. The proof in the multiblock case can be found in Robert and Casella (2004).

### 6.3.3 Metropolis Algorithm

A very general method of defining an MC is the Metropolis algorithm (Metropolis et al., 1953) as extended by Hastings (1970). This method differs from Gibbs sampling in that it can be used to simulate the entire vector  $\theta$  at each iteration of the algorithm. It is also used to sample full conditionals of a Gibbs sampler when they are unrecognizable. It is a remarkable algorithm that we show, for a special case, generates a Markov chain with the joint posterior as stationary distribution. Moreover, when used as a method of sampling unrecognizable full conditionals in Gibbs sampling, it can also be shown that a hybrid sampler that replaces a sample from a full conditional with one step of the Metropolis algorithm also has the joint posterior as its stationary distribution (see Robert and Casella, 2004). This algorithm is sometimes referred to as “Metropolis within Gibbs.”

The Metropolis algorithm is another type of accept-reject algorithm. It requires a *candidate generating distribution*; sometimes referred to as the *proposal distribution*. The algorithm begins with an initial value  $\theta^1$ . At the  $k$ th iteration, we have  $(\theta^1, \theta^2, \dots, \theta^k)$ . The  $k+1$ st iteration first generates  $\theta^*$  from a proposal density  $h(\theta^* | \theta^k)$ . This density should mimic the actual posterior

distribution in some sense, but in theory, it can be any distribution with the same support as the posterior. Define

$$\alpha(\theta^*, \theta^k) = \min \left\{ 1, \frac{p(\theta^*)h(\theta^k|\theta^*)}{p(\theta^k)h(\theta^*|\theta^k)} \right\} \equiv \alpha.$$

We then simulate  $U \sim U[0, 1]$  and we select  $\theta^{k+1} = \theta^*$  if  $U \leq \alpha$  and otherwise take  $\theta^{k+1} = \theta^k$ . Thus

$$\theta^{k+1} = \begin{cases} \theta^* & \text{with probability } \alpha(\theta^*, \theta^k) \\ \theta^k & \text{with probability } 1 - \alpha(\theta^*, \theta^k) \end{cases}.$$

Note that  $\alpha$  only uses the ratio of two values of  $p(\cdot)$ , so it is enough to know the kernel of the posterior density.

**EXAMPLE 6.3.3.** *Proposal Distribution for Poisson.* Suppose we have iid  $\text{Pois}(\theta)$  data with a conjugate gamma prior as in Subsection 5.3.1. A normal approximation with matching mean and variance is often a reasonable approximation to the Gamma posterior. If we didn't know how to sample from a  $\text{Gamma}(a, b)$  distribution, we could use the Metropolis algorithm with, say, a  $N(\theta^k, a/b^2)$  proposal distribution, namely,

$$h(\theta^*|\theta^k) \propto \exp\{-0.5(\theta^* - \theta^k)^2/(a/b^2)\}.$$

Note that  $h(\theta^*|\theta^k) = h(\theta^k|\theta^*)$ , so  $\alpha$  simplifies to

$$\alpha(\theta^*, \theta^k) = \min\{1, [\theta^*/\theta^k]^{a-1} \exp(b(\theta^k - \theta^*))\},$$

which is easy to compute.

**EXERCISE 6.13.** Write code in R to sample from a  $\text{Gamma}(a, b)$  distribution using the Metropolis algorithm with a normal proposal. You might scale the variance of the proposal to see if you can get the acceptance rate in the range of 20-40%. Plot your history of the chain and see if you can identify an appropriate burn-in value. Also try different initial values and compare histories. Try different values of  $(a, b)$ . For example, the normal candidate should work better if  $a$  is moderate to large.

The original Metropolis algorithm assumed that  $h(\theta^k|\theta^*) = h(\theta^*|\theta^k)$  so that  $\alpha(\theta^*, \theta^k) = \min\{1, p(\theta^*)/p(\theta^k)\}$ . This is called the *random walk*. In that case, it is easy to see that we use  $\theta^*$  if its density  $p(\theta^*)$  is larger than  $p(\theta^k)$ , the density for  $\theta^k$ . If the density  $p(\theta^*)$  for  $\theta^*$  is smaller than  $p(\theta^k)$ , we use  $\theta^*$  with probability  $\alpha(\theta^*, \theta^k)$ , which will be large when the density of  $\theta^*$  is nearly as big as the density for  $\theta^k$ , and will be small when the density of  $\theta^*$  is much smaller than the density for  $\theta^k$ . Later we will show that this gives the correct stationary distribution but now we discuss how to implement Metropolis.

Various suggestions have been made about how to choose  $h(\theta^*|\theta^k)$ . Often, it is taken as a  $N(\theta^k, \Sigma^k)$  distribution with various suggestions for  $\Sigma^k$ .

- I. Often one can find the posterior mode  $\theta_M$  of  $p(\theta)$  and, up to a constant multiple, the second derivative of  $\log[p(\theta)]$ , say,  $\ddot{\ell}(\theta)$ . Define

$$\tilde{\Sigma} = [-\ddot{\ell}(\theta_M)]^{-1}.$$

It is generally recommended to take

$$\Sigma^k = c\tilde{\Sigma},$$

where  $c$  is a “tuning” parameter chosen so that  $\theta^*$  values are accepted between 20% and 40% of the time.

- II. Another proposal is to pick any old  $\Sigma_0$ , use this on every iteration of the MC for a few thousand iterations, compute the sample covariance matrix of the  $\theta^k$ 's from this initial run, say,  $\tilde{\Sigma}$ , and use  $\Sigma^k = c\tilde{\Sigma}$  on all subsequent iterations, where again,  $c$  is a tuning parameter chosen to give an acceptance rate between 20% and 40%.
- III. A third proposal is similar to the second but uses only the diagonal elements of the sample covariance matrix.
- IV. Another proposal, called the *independence* proposal, simply samples from a fixed distribution with density, say  $g(\theta)$ , possibly a multivariate normal with mean vector equal to the posterior mode and with covariance selected in one of the above ways. Thus all of the proposed samples are iid from this density. However, since the acceptance probability depends on the last value in the chain, the resulting chain is indeed a Markov chain with dependence from one iteration to the next.

The actual Metropolis sampler is a mixture of continuous and discrete distributions since at each iteration, there is often positive probability of the new iterate being identically equal to the last iterate, and also positive probability that it will be the value that was taken from the (continuous) candidate generating distribution. If we use the Metropolis-within-Gibbs hybrid sampler, it is not immediately obvious that the overall chain has the appropriate stationary distribution. While it is not particularly difficult to establish this result, we refer the reader to Tierney (1994).

**EXERCISE 6.14.** (a) Write an algorithm to sample from the joint posterior for the  $Weib(\alpha, \lambda)$  model assuming independent Gamma priors for  $\alpha$  and  $\lambda$ , and using the Metropolis algorithm with a bivariate normal random walk proposal distribution. [This is a random walk in the sense that when at  $\theta^k$  the next proposed step to  $\theta^*$  consists of adding an independent  $N(0, \Sigma)$  variate to  $\theta^k$ .] (b) Write an algorithm to perform Gibbs sampling where the full conditional for  $\lambda$  is sampled directly and where one step of the Metropolis algorithm is used to sample the full conditional for  $\alpha$ . You must obtain an explicit formula for  $\alpha(\theta^*, \theta^k)$  in each case with appropriately defined  $\theta$ . Don't confuse the  $\alpha$  in the Weibull parametrization and the acceptance probability  $\alpha$  of the Metropolis algorithm! (c) Write R code to sample from these distributions. (d) Run your R code using the leukemia data of Example 6.3.2. Compare your results with the results from using WinBUGS.

### 6.3.3.1 Proof that $P(\theta)$ is the Stationary Distribution\*

Denote the MC as  $\theta^r$ ,  $r = 1, 2, \dots$ . This proof is for  $p(\theta)$  discrete. By changing  $(i, j, k)$  to  $(\theta^*, \theta^r, \theta^{r+1})$  and sums to integrals, the proof works in the “continuous” case. Unfortunately, there really isn’t a continuous case. The distributions  $p(\cdot)$  and  $h(\theta^* | \theta^r)$  may be continuous, but from the definition of the MC, the distribution of  $\theta^{r+1}$  given  $\theta^r$  is a mixture of a continuous distribution and a discrete distribution. This causes some technical (measure theoretic) difficulties. Simple justifications for the Metropolis algorithm gloss over this fact.

Given  $\theta^r = j$ , generate  $\theta^*$  from density  $h(i|j)$ . Define

$$\alpha(i, j) = \min \left\{ 1, \frac{p(i)h(j|i)}{p(j)h(i|j)} \right\}$$

and

$$\theta^{r+1} = \begin{cases} \theta^* & \text{with probability } \alpha(\theta^*, j) \\ j & \text{with probability } 1 - \alpha(\theta^*, j) \end{cases} .$$

We need to establish (3). Let  $g(k)$  be the density for  $\theta^{r+1}$ . We need to show that if the density for  $\theta^r$  is  $p(j)$  then  $g(k) = p(k)$ . Let

$$\delta_{ik} = \begin{cases} 1 & \text{if } i = k \\ 0 & \text{if } i \neq k \end{cases} .$$

From the definition of the Markov process, considering all the possibilities of  $(\theta^*, \theta^r) = (i, j)$

$$\begin{aligned} g(k) &= \sum_i \sum_j \delta_{ik} \alpha(i, j) h(i|j) p(j) + \sum_i \sum_j \delta_{jk} [1 - \alpha(i, j)] h(i|j) p(j) \\ &= \sum_j \alpha(k, j) h(k|j) p(j) + \sum_j \delta_{jk} \sum_i [1 - \alpha(i, j)] h(i|j) p(j). \end{aligned} \quad (8)$$

We look at each term individually, starting with the second

$$\begin{aligned} &\sum_j \delta_{jk} \sum_i [1 - \alpha(i, j)] h(i|j) p(j) \\ &= \sum_j \delta_{jk} \sum_{\{i: \alpha(i,j)=1\}} [1 - \alpha(i, j)] h(i|j) p(j) \\ &\quad + \sum_j \delta_{jk} \sum_{\{i: \alpha(i,j)\neq1\}} [1 - \alpha(i, j)] h(i|j) p(j) \\ &= 0 + \sum_j \delta_{jk} \sum_{\{i: \alpha(i,j)\neq1\}} \left[ 1 - \frac{p(i)h(j|i)}{p(j)h(i|j)} \right] h(i|j) p(j) \\ &= \sum_j \delta_{jk} \sum_{\{i: \alpha(i,j)\neq1\}} 1 h(i|j) p(j) - \sum_j \delta_{jk} \sum_{\{i: \alpha(i,j)\neq1\}} p(i)h(j|i) \\ &= \sum_j \delta_{jk} \sum_{\{i: \alpha(i,j)\neq1\}} [h(i|j) p(j) - p(i)h(j|i)] \\ &= \sum_{\{i: \alpha(i,k)\neq1\}} [h(i|k) p(k) - p(i)h(k|i)] \\ &= \sum_{\{j: \alpha(j,k)\neq1\}} [h(j|k) p(k) - p(j)h(k|j)]. \end{aligned}$$

Thus we have all together

$$g(k) = \sum_j \alpha(k, j) h(k|j) p(j) + \sum_{\{j: \alpha(j,k)\neq1\}} [h(j|k) p(k) - p(j)h(k|j)].$$

Now examine the first term:

$$\begin{aligned} &\sum_j \alpha(k, j) h(k|j) p(j) \\ &= \sum_{\{j: \alpha(k,j)=1\}} \alpha(k, j) h(k|j) p(j) + \sum_{\{j: \alpha(k,j)\neq1\}} \alpha(k, j) h(k|j) p(j) \\ &= \sum_{\{j: \alpha(k,j)=1\}} h(k|j) p(j) + \sum_{\{j: \alpha(k,j)\neq1\}} \frac{p(k)h(j|k)}{p(j)h(k|j)} h(k|j) p(j) \\ &= \sum_{\{j: \alpha(k,j)=1\}} h(k|j) p(j) + \sum_{\{j: \alpha(k,j)\neq1\}} p(k)h(j|k). \end{aligned}$$

So all together we have

$$\begin{aligned} g(k) &= \sum_{\{j: \alpha(k,j)=1\}} h(k|j) p(j) + \sum_{\{j: \alpha(k,j)\neq1\}} p(k)h(j|k) \\ &\quad + \sum_{\{j: \alpha(j,k)\neq1\}} [h(j|k) p(k) - p(j)h(k|j)]. \end{aligned}$$

Finally we note that by the definition of  $\alpha(k, j)$

$$\{j : \alpha(k, j) = 1\} = \{j : \alpha(j, k) \neq 1\} \cup \{j : p(k)h(j|k) = p(j)h(k|j)\} \quad (9)$$

so, using this decomposition and recalling that  $h(j|k)$  is a well-defined conditional density so that  $1 = \sum_j h(j|k)$ , we get

$$\begin{aligned}
g(k) &= \sum_{\{j: \alpha(j,k) \neq 1\}} h(k|j)p(j) + \sum_{\{j: p(k)h(j|k) = p(j)h(k|j)\}} h(k|j)p(j) \\
&\quad + \sum_{\{j: \alpha(k,j) \neq 1\}} p(k)h(j|k) + \sum_{\{j: \alpha(j,k) \neq 1\}} [h(j|k)p(k) - p(j)h(k|j)] \\
&= \sum_{\{j: \alpha(j,k) \neq 1\}} h(j|k)p(k) + \sum_{\{j: p(k)h(j|k) = p(j)h(k|j)\}} p(k)h(j|k) \\
&\quad + \sum_{\{j: \alpha(k,j) \neq 1\}} p(k)h(j|k) \\
&= \sum_j h(j|k)p(k) \\
&= p(k).
\end{aligned} \tag{10}$$

**EXERCISE 6.15.** Use the law of total probability and invent any necessary notation to establish why (8) holds. Also establish why (9) and (10) hold.

### 6.3.4 Slice Sampling

Slice sampling uses a Markov chain to sample from a *univariate* distribution with density  $p(\theta)$ . Since we generally won't know the constant of integration, we work with the kernel of the posterior,  $p_*(\theta)$ . Imagine the two-dimensional graph of  $p_*(\theta)$ . Slice sampling takes a random walk through the area under the graph of  $p_*(\theta)$ , alternating steps along the orthogonal axes. To sample a slice of this method

- I. Choose an arbitrary  $\theta^1$  in the support of  $p_*(\theta)$  to initialize the chain. The random walk starts at  $(\theta^1, 0)$ .
- II. Given  $\theta^1$ , step up towards the density by simulating  $u^1 \sim U[0, p_*(\theta^1)]$  and move to  $(\theta^1, u^1)$ .
- III. From a point  $(\theta^k, u^k)$  with a *unimodal*  $p_*(\theta)$ , step along the  $\theta$  axis by finding the two points  $\theta_{k1} < \theta_{k2}$  with  $u^k = p_*(\theta_{k1}) = p_*(\theta_{k2})$  and simulate  $\theta^{k+1} \sim U[\theta_{k1}, \theta_{k2}]$ . Move to  $(\theta^{k+1}, u^k)$ . For non-unimodal densities, find  $\{\theta : p_*(\theta) \geq u^k\}$  and sample from a uniform distribution on that set, say  $U[\theta : p_*(\theta) \geq u^k]$ . The main difficulty in slice sampling is finding the set  $\{\theta : p_*(\theta) \geq u\}$  when  $p_*(\theta)$  is not unimodal.
- IV. Simulate  $u^{k+1} \sim U[0, p_*(\theta^{k+1})]$  and move to  $(\theta^{k+1}, u^{k+1})$ .

The random walk used for slice sampling is a special case of Gibbs sampling since we alternate between sampling  $U|\theta = \theta^* \sim U[0, p_*(\theta^*)]$  and  $\theta|U = u \sim U[\theta' : p_*(\theta') \geq u]$ .

It is not difficult to prove that  $p$  is the stationary distribution of the subchain  $\theta^k$ , and that it is also aperiodic. See Robert and Casella (2004) for details.

**EXERCISE 6.16.** Let  $y \sim \text{Exp}(\theta)$ . Give an explicit algorithm for sampling from this distribution using the slice sampler.

**EXERCISE 6.17.** Assume that there exists a joint density that gives rise to the full conditionals for the slice sampler, i.e., assume that  $p(\theta, u)$  gives rise to  $p_{\theta|u}(\theta) = p(\theta, u)/p_u(u)$  and  $p_{u|\theta}(u) = p(\theta, u)/p_\theta(\theta)$ . It immediately follows that  $p_{\theta|u}(\theta)/p_{u|\theta}(u) = p_\theta(\theta)/p_u(u) \propto p_\theta(\theta)$ . Thus under this assumption, which can be verified for this problem, obtain the kernel of the pdf for  $\theta$ .

**EXERCISE 6.18.** We say that the full conditionals for the slice sampler are *compatible* if they uniquely determine a joint distribution. A necessary and sufficient condition for the existence of a

joint density that corresponds to two given conditional densities is (i) that the support of the two full conditionals are identical, i.e.,  $\{(\theta, u) : p(\theta|u) > 0\} = \{(\theta, u) : p(u|\theta) > 0\}$  and (ii) the ratio of kernels for the full conditionals, say the kernel for  $\theta|u$  divided by the kernel for  $u|\theta$ , is integrable as a function of  $\theta$  (Arnold and Press, 1989). (a) Using this result, establish that the joint density for  $(\theta, u)$  for the slice sampler exists. (b) Consider the two conditional distributions:  $x|y \sim \text{Exp}(y)$  and  $y|x \sim \text{Exp}(x)$ . Applying the Arnold and Press result, establish that there does not exist a joint density that is compatible with these two full conditionals.

### 6.3.5 Checking MCMC Samples

Our first check on convergence of a Markov chain to the stationary distribution  $p(\theta)$  is to plot histories  $(k, \theta_j^k)$  for each component of the parameter vector  $j = 1, \dots, r$ , as shown in Figure 6.1. After a burn-in, the chain should not show any trends or patterns. An essentially ideal plot looks like Figure 6.1(a) and the worst plots look like Figure 6.1(c).

A second diagnostic is the autocorrelation function. If there is a lot of autocorrelation, we may have to sample a huge number of values to get reasonable numerical accuracy for our inferences. See Figure 6.2.

WinBUGS provides an estimate of Monte Carlo error that is analogous to a frequentist standard error. For example, if  $\theta^1, \dots, \theta^m$  are iid from  $p(\theta)$ , a component of the vector  $\theta$ , say  $\theta_j$ , has a posterior mean estimated by the sample mean  $\bar{\theta}_j$  and a posterior variance estimated by the sample variance  $s_j^2$ . The estimated standard deviation of  $\bar{\theta}_j$  is  $[s_j^2/m]^{1/2}$  and an approximate 95% confidence interval for the posterior mean has endpoints  $\bar{\theta}_j \pm 2[s_j^2/m]^{1/2}$ . If the sample size  $m$  is large enough so that the confidence interval is, say,  $10 \pm 0.0001$ , we will be quite happy with the Monte Carlo approximation. On the other hand if our interval is  $10 \pm 0.5$  or  $0.37 \pm 0.1$ , we might want to increase  $m$ . WinBUGS provides an estimated standard deviation for  $\bar{\theta}_j$  that is appropriate for the dependent samples obtained from the Markov chain.

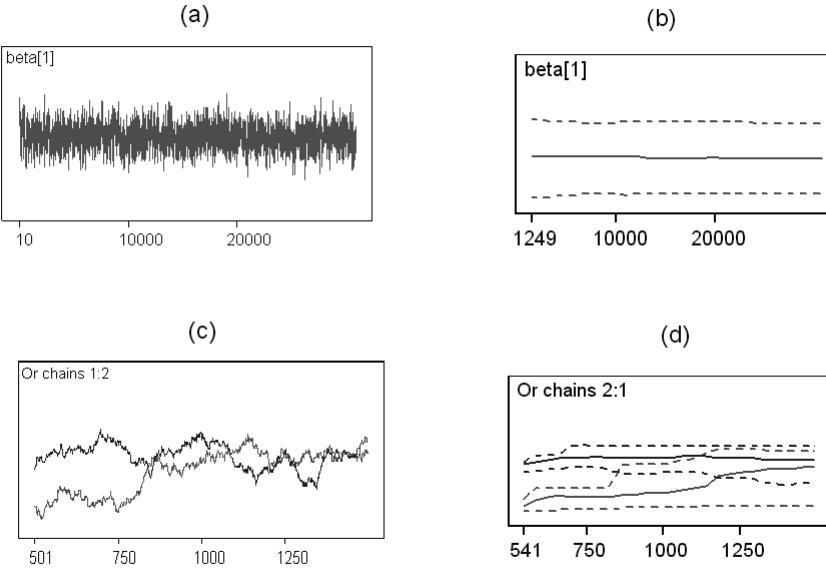


Figure 6.3: History and quantile plots for chains that: (a–b) have converged after 10 iterations, and (c–d) chains that are not converging.

Another plot of interest looks at running quantiles as the Markov chain is sampled. It is not easy to estimate the 0.025 and 0.975 quantiles of any sample, since there are fewer data in the tails of the distribution. What we look for in quantile plots is stability after a burn-in phase. Ideally, plots of quantiles are constant through time. Figure 6.3 gives two quantile plots. Figure 6.3(a) shows a Markov chain history that has converged with a burn-in of only 10 iterations and the corresponding quantile plots are in (b). The quantiles plotted are the median, and the 0.025 and 0.975 quantiles, calculated using samples up to the current iteration, across all iterations. Figure 6.3(c) shows the same chains (with two distinct starting values) that were given in Figure 6.1(c); the corresponding quantile plots are in Figure 6.3(d). The quantile plots show two running medians, and two running 0.025 and 0.975 quantiles, which are obviously running amok, indicating serious lack of convergence.

Given a burned-in, thinned chain, one could perform control charts on the data. Control charts for means are specifically designed to validate the assumption that a sample consists of iid observations, cf. Christensen (2001b).

Perhaps the best check is to see whether multiple chains that are initialized at points spread out in the parameter space ultimately converge to the same distribution. Figure 6.1 illustrated convergence and nonconvergence of chains.

A more formal approach takes the following course. Sample independent initial values,  $\theta_1^1, \dots, \theta_m^1$  from a highly dispersed initial distribution  $q(\theta)$ . For each  $i = 1, \dots, m$ , independently generate chains  $\theta_i^1, \theta_i^2, \dots, \theta_i^m$ . After a burn-in of  $BI$  observations and thinning so that the observations are approximately independent, there should remain, say,  $s$  observations on each of  $m$  groups. This constitutes data for a balanced one-way ANOVA. Moreover, if the chains have converged to the same distribution, the means should be the same in each group, so the null hypothesis for the standard analysis of variance  $F$  test should hold. We can also compare the sample variances and other characteristics of the multiple chains. Gelman and Rubin (1992) and Brooks and Gelman (1998) provide related diagnostics.

**EXERCISE 6.19.** Using the Weibull model, the data in Example 6.3.2, and the prior specified in Exercise 6.14, run the model with three distinct initial values for both parameters and look at the quantiles plot, history plots, and autocorrelation plots. Determine an appropriate value of the burn-in. Try thinning the chains to see if you can reduce any autocorrelation. Pay particular attention to the Monte Carlo error that is reported and make sure that you ultimately have a large enough MC sample size to achieve reasonable accuracy for all of the posterior means. Comment on all of these issues in a short report.

---

## Chapter 7

---

# Basic Concepts of Regression

---

### 7.1 Introduction

This chapter explores models for predicting a response variable based on one or more predictor variables (covariates). The process is commonly known as regression analysis although originally that term was restricted to the prediction of a continuous measurement variable by a collection of other continuous measurement variables.

An appropriate regression model for statistical analysis depends on the nature of the response variable. Chapter 8 presents methods for analyzing binary responses (cancer, no cancer) and binomial responses (number of cures out of a fixed number of patients). Chapter 9 examines standard measurement data (lengths, weights) using normal distributions. Subsequent chapters consider counts that have no obvious upper limit (armadillos caught on a hunting trip, flaws on a DVD) using Poisson and negative binomial distributions, as well as time to event data (time until your car breaks down, survival after a cancer diagnosis) using exponential, Weibull, and log-normal distributions. In all cases, the predictor variables (covariates) can be continuous measurements (traditional regression models), categorical variables (factors) used to indicate group status (traditional analysis of variance models), or a combination of these (traditional analysis of covariance models).

We focus primarily on models that involve linear combinations of the predictor variables, i.e., each predictor variable is multiplied by an unknown parameter called a *regression coefficient* and they are added together. Depending on the nature of the response variable, different functions of the linear combinations are considered. Traditionally the regression coefficients are all treated as fixed numbers but increasingly *mixed effects models* are being used in which some regression coefficients are treated as random variables. Ultimately, a Bayesian analysis treats all the regression coefficients as random, but priors are specified differently for mixed model regression coefficients that are random in the sampling model.

Regression data can arise from almost any experimental or observational study design. Commonly used experimental designs include completely randomized designs, randomized block designs, and split-plot/nested/hierarchical/multi-level designs. Experimental designs often incorporate factorial treatment structures which constitute a very efficient way of defining experimental treatments and that allow the consideration of many factors in one experiment. The most widely accepted experimental designs use random allocation of subjects into treatment groups to provide a philosophical basis for concluding that the treatments actually cause any observed experimental effects. In medicine, randomized experiments are known as *clinical trials*.

Other data collection schemes are collectively known as observational studies. These include *cohort* or *prospective* studies in which a sample (ideally a random sample) is collected from a population and various data are collected on each individual including the value of the response variable. Sometimes multiple response variables are measured over time. When restricted to observations taken at a single point in time, these studies are also known as *cross-sectional*. Most statistical results are developed for experiments and prospective data.

Sometimes, prospective studies are impractical and one must employ *case-control* studies (also called *retrospective* studies). For example, when studying a rare disease such as ALS (Lou Gehrig's disease), a cross-section of the population would have to be huge to include a reasonable number of

people with the disease. In such instances we might take two samples, one from people with the disease (cases) and one from people without the disease (controls). We then seek to find characteristics of the people that can distinguish or *discriminate* between the two populations.

Regression methods are developed for prospective studies. They are easily adapted to experimental situations in which the predictor variables are fixed, not random. With suitable modifications, they can often be used to analyze discrimination data, but naive use of regression methods for discrimination can lead to inappropriate conclusions.

Regression models typically address one or more of three scientific aims. They can provide point or interval predictions for future observations given particular values of the predictor variables. (What is the probability of having a heart attack in the next 10 years given your age, height, weight, cholesterol, and blood pressure?) Second, from the collection of all predictor variables, they can identify subsets that are most useful for prediction. (Do height and weight really help the prediction process?) Third, we can quantify the association between the response and one or more predictor variables of interest while controlling for the influence of other variables. (Given the other predictor variables including systolic blood pressure, how much does diastolic blood pressure improve our predictions?)

This chapter presents an introduction to concepts that are common to a variety of regression applications. In particular, it explores the use of linear combinations of predictor variables to address issues of interest in regression analysis. This includes handling binary and multi-category predictors, confounding, and interactions. Many concepts and methods are illustrated using a case study that investigates the association between lung function and smoking in adolescents.

**EXAMPLE 7.1.1. FEV Data.** Rosner (2006) presents data on pulmonary function (lung capacity) in adolescents. The response is forced expiratory volume (FEV) which measures the volume of air in liters expelled in 1 second of a forceful breath. Lung function is expected to increase during adolescence, but smoking may slow its progression. The association between smoking and lung capacity in adolescents is investigated using data from 345 adolescents between the ages of 10 and 19. The predictor variables include age and smoking status (NS-not current smoker and S-current smoker).

We want to compare mean FEV for smokers and nonsmokers. A simple two-sample normal analysis as in Chapter 5 is misleading because smokers have higher FEVs. This occurs because high FEVs are strongly associated with age whereas smoking is also much more common among older subjects. An appropriate analysis of the effect of smoking must account for age.

Although our case study involves linear regression, the methods used are common to all regression methodologies. A thorough treatment of linear regression is left to Chapter 9. In this chapter, Section 2 defines notation and Section 3 discusses regression modeling when the goal is prediction. Section 4 develops the nuts-and-bolts details of regression modeling with linear structures, including the concepts of confounding and effect modification (interaction); Section 5 provides an illustration.

## 7.2 Data Notation and Format

Let  $y$  denote a random response variable and  $x = (x_1, x_2, \dots, x_r)'$  a vector of predictor variables also called the *covariate combination*. *Simple regression problems have  $r = 2$  with  $x_1 \equiv 1$ .* In such cases it is often convenient to write  $x \equiv x_2$ , the only nontrivial predictor variable. In applications it is common to use descriptive terms for regression variables. For example, the response  $y$  may be FEV with predictors  $x_1 \equiv 1$  a constant;  $x_2 = \text{Smoker}$ , a binary variable that is 1 if the person is a smoker and 0 otherwise; and  $x_3 = \text{Age}$ , a (theoretically) continuous variable. Nearly all models that do not incorporate categorical variables involve a constant variable.

One key aspect of a categorical (factor) variable is that with, say four categories, it must be rewritten as a corresponding collection of four binary  $x_j$  variables wherein  $x_j = 1$  is used to indicate

that the individual is in one particular category. Thus, a categorical variable that identifies the four high school classes—senior, junior, sophomore, freshman—is rewritten as four binary variables  $x_1, \dots, x_4$  where each binary variable is used to identify one of the four classes. As discussed later, in models with a constant, we need only include three of the four binary  $x_j$ s. It does not really matter which of the four we delete or whether we delete the constant instead, but the interpretations of the regression coefficients depend on which variable is deleted.

In the FEV example with  $x_1 \equiv 1$ ,  $x_2 = \text{Smoker}$ , and  $x_3 = \text{Age}$ , we should normally define both  $x_2$  and a second binary variable, say  $x_4 = \text{Nonsmoker}$ . Since we planned to delete  $x_4$ , we did not bother to define it. Nonetheless, our analysis would be equivalent using predictor variables: constant, Age, Nonsmoker; or using the variables: Age, Smoker, Nonsmoker. Many computer programs facilitate this process of defining indicator variables by allowing the user to specify which variables are factors and which are continuous covariates. They then (internally) create the binary  $x_j$  variables and automatically eliminate a variable (sometimes further modifying the  $x_j$ s). Many programs assume the variables are one type (continuous or categorical) and the user must specify which variables are of the other type. We will not be using such programs so categorical variables will need to be manually transformed into appropriate  $x_j$  variables. More details on categorical predictors are presented in the next section.

Table 7.1 presents a general form for a regression data file. The typical format of regression data is to list the information for each individual in a single row, i.e., to use one row for each data record. Data files may or may not contain an initial row of variable names. The columns can be permuted to any order. In general the data elements are written  $y_i$  and  $x_{ij}$  where  $i$  indicates the subject and  $j$  indicates a particular predictor variable. The vector of all predictor variables for subject  $i$  is written  $x_i$ . It should (we hope) be clear from the context whether, say,  $x_2$  means the second predictor variable or the vector of predictors associated with the second individual.

Table 7.1: *Example of a general regression data format.*

Subject	$y$	$x_1$	$x_2$	...	$x_r$
1	$y_1$	$x_{11}$	$x_{12}$	...	$x_{1r}$
2	$y_2$	$x_{21}$	$x_{22}$	...	$x_{2r}$
3	$y_3$	$x_{31}$	$x_{32}$	...	$x_{3r}$
:	:	:	:		:
$n$	$y_n$	$x_{n1}$	$x_{n2}$	...	$x_{nr}$

EXERCISE 7.1. Read the FEV data into WinBUGS. A Microsoft Excel data file is available at the book website. The FEV data file contains 345 records. Table 7.2 presents some of the FEV data in WinBUGS format. Key features of the WinBUGS format are that left and right brackets follow each variable name and the word END appears after the last data record.

Table 7.2: *WinBUGS format of the FEV data file.*

Age[ ]	FEV[ ]	Smoke[ ]
19	5.102	0
18	4.22	0
18	4.086	1
:	:	:
10	2.1	0
END		

Databases are frequently created in Microsoft Excel. A simple way to transfer data is to copy and paste from Excel into WinBUGS. Unfortunately this can result in a mess of data that is one long uninterrupted sequence in WinBUGS. Try it yourself. Copy the header and the first 3 FEV data lines (or all the lines) in Excel and paste them into WinBUGS. Your data might appear in WinBUGS as

AgeFEVSmoke195.1020184.220184.0861

One “high tech” method for getting around this is to copy the data from Excel and paste it into a text editor like Notepad. Then copy the data from Notepad and paste it into WinBUGS.

### 7.3 Predictive Models: An Overview

Consider taking a random sample from a population. They might be students from the University of Kentucky, millionaires from Orange County, fishermen from Minnesota, or rugby players from New Zealand. Some of the different response variables  $y$  that we might consider are: whether cholesterol is over 240 (binary), total cholesterol (measurement), number of mosquito bites in the last week (count), or time since last traffic ticket (time to event). Our predictor variables  $x_j$  are limited only by our imagination and ability to measure/count. Clearly, both the responses and the predictors are random, so the population determines a joint distribution for  $x$  and  $y$ .

If we know the joint distribution, an optimal predictor of  $y$  is the conditional expected value  $E(y|x)$ , see Christensen (2002, Sec. 6.3). This *best predictor* is the mean of the response given the covariate combination. In the case of binary (Bernoulli) data, the mean is the probability of success given the covariate combination. To simplify notation, write the conditional mean as

$$m(x) \equiv E(y|x).$$

The best predictor is also known as the *regression function*.

If we know the joint distribution of  $x$  and  $y$ , we know  $m(x)$ . Our problem is that we do not typically know the joint distribution, so we need to estimate  $m(x)$  from data. R. A. Fisher recognized that in regression analysis, the distribution of  $x$  is of little importance and that the key features are the regression function and the conditional distribution of  $y$  given  $x$ . Thus we commonly treat  $x$  as a fixed quantity. It also follows that regression models apply not only to prospective studies with random  $x$  but also to designed experiments in which  $x$  is fixed by the investigator, see also Aldrich (2005).

There are several strategies for modeling the regression function  $m(x)$ . Mathematically, the simplest common strategy is to assume a *linear regression (multiple regression)* model

$$m(x) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_r x_r = x' \beta,$$

where  $\beta = (\beta_1, \beta_2, \beta_3, \dots, \beta_r)'$  is a vector of unknown regression parameters (coefficients).

While linear regression frequently works well for measurement data, the model is ridiculous for binary data. For binary data, the mean is a probability so  $0 \leq m(x) \leq 1$ , whereas the multiple regression has no such restriction. *Generalized linear models* assume that

$$m(x) = g(x' \beta)$$

for some known function  $g$  but unknown  $\beta$ . For binary data, the most common choice for  $g$  is the *logistic* function

$$g(x' \beta) = \frac{e^{x' \beta}}{1 + e^{x' \beta}}$$

but other transformations that make the numbers fall between 0 and 1 are also sometimes used. For time-to-event data and especially for count data, the most common choice is probably  $g(x' \beta) = \exp(x' \beta)$ . In general, appropriate choices for  $g$  depend on the nature of the response variable.

It is common practice to write generalized linear models using the inverse function of  $g$  as

$$g^{-1}[m(x)] = x'\beta$$

wherein the function  $g^{-1}$  is called the *link function*. For  $g(x'\beta) = \exp(x'\beta)$ , we get

$$\log[m(x)] = x'\beta,$$

which is called a *log-linear model*. When

$$m(x) \equiv p(x)$$

is a probability and  $g$  is the logistic function, then the inverse of  $g$  is the log odds of the probability, also known as the *logit* function. The logit model is

$$\text{logit}[p(x)] \equiv \log\left[\frac{p(x)}{1-p(x)}\right] = x'\beta.$$

To analyze any generalized linear model one needs to use both  $g$  and its inverse. The procedure can be named after either function so that logit models and logistic regression models are the same thing.

An even more general model than the generalized linear model involves a known function  $h(\cdot)$  but relaxes the requirement that the predictors and regression coefficients combine linearly. An example is

$$m(x) = \beta_1 + \beta_2 x_2 + e^{\beta_3 x_3}.$$

In general we write

$$m(x) = h(x; \beta).$$

When applied to measurement data with normal distributions, this model is known as *nonlinear regression*. With a Bernoulli response we might apply the logistic function to a function  $h(x; \beta)$  and call this nonlinear logistic regression or logistic nonlinear regression. Nonlinear regression is discussed briefly in Section 9.8.

The most general model is almost no model at all. If  $m(x)$  is continuous and we have a great deal of data, we can just average the  $y$  data collected near  $x$  to estimate  $m(x)$ . The trick is deciding what “near” means and having enough data. Alternatively, mathematical results establish that for sequences of known (basis) functions  $\phi_j(\cdot)$ ,

$$m(x) = \sum_{j=1}^{\infty} \beta_j \phi_j(x).$$

Here we just need to estimate the unknown  $\beta_j$ s. In simple regression the  $\phi_j$  functions are often polynomials  $\phi_j(x) = x^{j-1}$  or cosines  $\phi_j(x) = \cos(2\pi j x)$  [ $x$  needs to be normalized first], or sines and cosines, or wavelets. In practice, the infinite sum must be replaced by a finite sum, the order of which needs to be determined. These two approaches are examples of *nonparametric regression*. Note that the basis function approach to a simple regression corresponds to a multiple regression

$$m(x) = \beta_1 \phi_1(x) + \cdots + \beta_r \phi_r(x) = [\phi_1(x), \dots, \phi_r(x)]\beta.$$

The basis function approach to nonparametric regression is discussed in more detail in Chapter 15.

The process of defining interesting classes of functions for  $m(x)$  and estimating an appropriate member of that class to use as a predictor is limited only by the imagination and ingenuity of the people analyzing the data. For example, something as complicated as neural networks fits easily within the framework of this discussion, cf. Hastie et al. (2001, Chapter 11).

The key features of all prediction models are the assumed distribution of  $y$  given  $x$  and the assumed form of  $m(x)$ . A further typical assumption is that observed data  $y_i$  are independent. Any

good statistical analysis does as much as practically possible to validate all of these assumptions. In particular, Christensen (2007) has shown that regardless of the conditional distribution, when we have the correct regression function  $m(x)$  and we look at any function of the predictor variables  $f(x)$ , we get  $\text{Cov}[y - m(x), f(x)] = 0$ . Thus, after specifying a model for  $m$  and then estimating  $m$  with  $\hat{m}$ , plots of residuals  $y_i - \hat{m}(x_i)$  versus any single variable  $x_{ij}$  or any function  $f(x_i)$  should display near zero correlation. Any violation of this result suggests that  $\hat{m}$  is a poor substitute for the true regression function. Perhaps we picked a poor model for  $m(x)$ . Moreover, the zero covariance result holds regardless of whether  $x_{ij}$  was included in the process of modeling  $m$ . For example, a quadratic trend in the residual versus  $x_{ij}$  plot indicates a correlation between the residuals and the (possibly mean adjusted) variable  $x_{ij}^2$ . We then need to modify  $\hat{m}$ . The most obvious idea is to add an unknown multiple of  $x_{ij}^2$  to our current model for  $m(x)$ .

#### 7.4 Modeling with Linear Structures

This section presents methods for working with linear combinations of the predictor variables. As applied in linear models and generalized linear models we discuss continuous predictors, binary predictors, using a single multi-category predictor, and the use of multiple multi-category predictors. The important issues of confounding and effect modification (interaction) are also discussed.

##### 7.4.1 Continuous Predictors

Continuous predictors are probably the most straightforward to use. For example, a simple linear regression has the form  $m(x) = \beta_1 + \beta_2 x$ . However, the predictive ability of models may be improved by transforming individual predictor variables. The simple linear regression may not fit the data as well as, say,  $m(x) = \beta_1 + \beta_2 x^2$ . Similarly, an untransformed simple logistic regression  $\text{logit}[m(x)] = \beta_1 + \beta_2 x$  may not fit as well as  $\text{logit}[m(x)] = \beta_1 + \beta_2 \log(x)$ . The point of transforming predictor variables is to improve the validity of model assumptions, i.e., get a better model for  $m(x)$ .

Regression models can include any transformation of any of the continuous original predictor variables. Log transformations are often effective when the minimum value of the predictor is an order of magnitude smaller than the maximum value. Power transformations  $x^\lambda$  are also often employed, see Christensen (1996, Section 7.10). Moreover, entire sequences of transformations of a single predictor can be used, e.g., the basis function approach to nonparametric regression. New predictors can also be constructed as functions of several of the original predictors. In multiple regression, a simple example creates a new predictor that is the product of two original predictors, see also Subsection 7.4.7.

##### 7.4.2 Binary Predictors

It is common in multi-predictor regression analysis to have some covariates that are continuous and some that are categorical. The simplest categorical predictors are binary. Binary variables classify subjects into one of two distinct groups such as sex (male/female) or disease exposure status (exposed/unexposed).

When used mathematically in predictive models all categorical variables, including binary predictors, use a 0/1 coding scheme. For example, in the FEV study, smoking status was coded with a 1 for smokers and a 0 for nonsmokers. These 0/1 variables are called *indicator variables* (sometimes also called *dummy variables*) with the “1” indicating inclusion in the labeled group. Although mathematically the 0/1 coding scheme is required, most computer programs are flexible enough to allow any two distinct symbols to be used as codes for the groups.

Let  $x_1 \equiv 1$  and  $x_2$  be an indicator for, say, smoking. Let  $y$  be continuous like FEV. A frequent goal of statistical analysis is to compare mean response for the two groups (smoking and non). A

simple linear regression in this setting models mean response as

$$m(x) = \beta_1 + \beta_2 x_2 = \begin{cases} \beta_1, & \text{group 1} \\ \beta_1 + \beta_2, & \text{group 2} \end{cases}. \quad (1)$$

The mean response for people in group 1 (nonsmoking) is given by  $\beta_1$ , and the mean response for group 2 (smoking) is  $\beta_1 + \beta_2$ . Therefore,  $\beta_2$  represents the change in the mean response from group 1 to group 2, i.e., the effect of being a member of the smoking group. Note that this is distinct from the effect of smoking, see Subsection 7.4.6.

Similarly, if our dependent variable  $y$  was an indicator of lung cancer, we might fit a simple logistic regression

$$\text{logit}[m(x)] = \beta_1 x_1 + \beta_2 x_2 = \beta_1 + \beta_2 x_2 = \begin{cases} \beta_1, & \text{group 1} \\ \beta_1 + \beta_2, & \text{group 2} \end{cases}. \quad (2)$$

The interpretation of the regression coefficients is similar but is for functions of the group means.  $\beta_1$  is a log odds and  $\beta_2$  is a log odds ratio, see Exercise 7.2.

Now consider a slightly different approach to defining the predictors. Let  $x_1$  be the indicator of smoking and let  $x_2 \equiv 1 - x_1$  be the indicator for nonsmoking. In this case, the linear regression becomes

$$m(x) = \beta_1 x_1 + \beta_2 x_2 = \begin{cases} \beta_1, & \text{group 1} \\ \beta_2, & \text{group 2} \end{cases}. \quad (3)$$

Here  $\beta_1$  is the group mean for smokers and  $\beta_2$  is the group mean for nonsmokers. In this model, the effect of being in the smoking group is  $\beta_2 - \beta_1$ . For all practical purposes, this model is equivalent to model (1). Only the interpretations of the regression coefficients have changed. The estimated means for each group and the estimated difference between groups will be identical. However, when doing a Bayesian analysis, the prior distribution on a parameter depends on the interpretation of the parameter, so understanding the interpretation is important.

**EXERCISE 7.2.** *Logistic Regression.* Consider the response variable disease presence ( $D = 1$ ) or absence ( $D = 0$ ) with a binary exposure variable ( $E = 1$  if exposed and  $E = 0$  if unexposed). Under a logistic regression model like (2), derive formulas for the following parameters: the odds of having disease for someone with exposure; the odds of having disease for someone without exposure; the odds ratio of disease for exposed versus unexposed; the probability of disease for exposed individuals; the probability of disease for unexposed individuals; the probability ratio (also called relative risk or risk ratio) of disease for exposed relative to unexposed. Explain how to obtain a Bayesian estimate of the probability ratio.

#### 7.4.3 Multi-Category Predictors

The methods presented in the previous subsection for binary covariates extend to categorical predictors (factors) that have 3 or more levels. With multi-category predictors, mathematically we need an indicator variable for each category. Similar to the previous subsection, for models that include a constant variable, one of the indicators can be eliminated. Alternatively, there is no strong reason to include a constant variable when categorical predictors are used.

Categorical covariates with multiple levels include high school class (freshman, sophomore, junior, senior), highest educational level (< high school, high school degree, college degree), and race. Categorical variables can be created by stratifying continuous variables according to cutpoints. For example, age can be categorized as (Young, Middle, Old). The number of levels in a categorical covariate will be denoted  $K$ . For example, the covariate sex is binary so it has  $K = 2$  levels, while high school class has  $K = 4$  levels.

Consider a linear model based on the factor high school class. We use an intercept and indicators  $x_2, x_3, x_4$  for junior, sophomore, freshman, which makes senior a *baseline* or *reference* category.

The model is

$$y|x, \beta, \tau \sim N[m(x), 1/\tau]$$

$$m(x) = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 = \begin{cases} \beta_1, & \text{senior} \\ \beta_1 + \beta_2, & \text{junior} \\ \beta_1 + \beta_3, & \text{sophomore} \\ \beta_1 + \beta_4, & \text{freshman} \end{cases}. \quad (4)$$

The first line of the display defines the conditional distribution and the second defines the model for  $m(x)$ . Together this model is (one version of) a one-factor *analysis of variance (ANOVA)* model for  $K = 4$  groups.

While  $\beta_1$  is the mean for the reference group (senior), all of the other regression coefficients are changes in the mean of the indicated group relative to the reference group. For group 2 and above, the mean of group  $k$  is  $\beta_1 + \beta_k$  so  $\beta_k = (\beta_1 + \beta_k) - \beta_1$  is the differential effect between group  $k$  and the baseline. Rather than focusing on  $\beta_k$ , we could look at other measures of group effects, for example the mean relative to the baseline,  $1 + \beta_k/\beta_1$ . Posterior probability intervals and probabilities of positive effects, i.e.,  $\Pr(\beta_k > 0|y)$ , for each  $\beta_k$  provide useful information for comparison to the baseline group. We can also easily compare any two groups. For instance, the difference in mean response between groups 2 and 4 is  $(\beta_1 + \beta_2) - (\beta_1 + \beta_4) = \beta_2 - \beta_4$ .

For binary responses  $y$  we can use the same predictors to fit a *logistic one-way ANOVA* model

$$y|x, \beta \sim \text{Bern}[m(x)]$$

$$\text{logit}[m(x)] = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 = \begin{cases} \beta_1, & \text{senior} \\ \beta_1 + \beta_2, & \text{junior} \\ \beta_1 + \beta_3, & \text{sophomore} \\ \beta_1 + \beta_4, & \text{freshman} \end{cases}.$$

The interpretation of regression coefficients is similar but involves log odds and odds ratios, see Exercise 7.3.

**EXERCISE 7.3.** Consider a logistic regression similar to Exercise 7.2 but now the exposure variable has 3 levels (no exposure, low exposure, high exposure). The response variable is disease status ( $D = 1$  or  $D = 0$ ).

- (a) Write down the logistic regression model when the unexposed group is the baseline group.
- (b) Derive formulas for the probability of disease in each group.
- (c) Derive formulas for the odds of disease in each group.
- (d) Derive formulas for odds ratios and risk ratios comparing the 3 groups. There should be 3 odds ratio formulas and 3 risk ratio formulas.
- (e) Suppose you have  $m$  simulated values from the posterior distribution of  $(\beta_1, \beta_2, \beta_3)$ . Explain how to obtain Bayesian estimates of the parameters in parts b, c, d.

In general, this approach to dealing with a categorical variable having  $K$  levels is to let  $x_1 \equiv 1$ , then pick a *reference (baseline)* category and create  $K - 1$  indicator variables,  $x_2, \dots, x_K$ , one for each category other than the baseline. Thus,  $x_k = 1$  for those individuals who belong to category  $k$ , and  $x_k = 0$  for those who belong to any of the other categories. Each individual  $i$  who is not in the reference group will belong to exactly one category of the  $K - 1$  variables, so will have only one value  $k$  with  $x_{ik} = 1$  among  $k = 2, \dots, K$ . Individuals in the reference group have  $x_{ik} = 0$  for all  $k = 2, \dots, K$ . We get to pick the baseline group as whatever we want. Typically it is the control group if one exists.

Just as models (1) and (3) are equivalent for two groups, an equivalent alternative to model (4) drops the intercept and uses  $K = 4$  indicator variables,  $x_1, \dots, x_4$ , one for each category. In other

words, model (4) involves a variable  $x_1$  that always equals 1 whereas an equivalent model uses a variable  $x_1$  that is an indicator of senior. The other variables remain the same. The one-way ANOVA becomes

$$y|x, \beta, \tau \sim N[m(x), 1/\tau]$$

$$m(x) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 = \begin{cases} \beta_1, & \text{senior} \\ \beta_2, & \text{junior} \\ \beta_3, & \text{sophomore} \\ \beta_4, & \text{freshman} \end{cases}. \quad (5)$$

The difference in means between groups 1 and  $k$  is now  $\beta_k - \beta_1$  and the relative mean is  $\beta_k/\beta_1$ .

Linear models that incorporate one or more categorical variables along with one or more continuous variables are traditionally called *analysis of covariance* (ACOVA or ANCOVA) models. Consider adding the continuous variable age,  $x_5$ , to our one-way ANOVA model (4) based on high school class, to get the one-way ACOVA

$$m(x) = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5. \quad (6)$$

The ACOVA model fits a separate simple linear regression line in age,  $x_5$ , to each group but the regression lines all have the same slope (are parallel). For the baseline group the line is  $m(x) = \beta_1 + \beta_5 x_5$  and for any other group  $k$  the line is  $m(x) = (\beta_1 + \beta_k) + \beta_5 x_5$ , in which the intercept of the line is  $(\beta_1 + \beta_k)$ . The slope for every line is  $\beta_5$ . Note that for each non-baseline group  $k$ ,  $\beta_k$  still gives the mean difference between that group and the baseline.

The ACOVA model (6) extends the ANOVA model (4). Alternatively, we could extend ANOVA model (5) to an ACOVA by adding the continuous variable  $x_5$ :

$$m(x) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5. \quad (7)$$

Now the simple linear regression line in  $x_5$  for group  $k$  is  $m(x) = \beta_k + \beta_5 x_5$ . Once again we have parallel lines for each group, we have merely renamed the intercept. Many people prefer the parameterizations of the one-way ANOVA model (5) and ACOVA model (7).

#### 7.4.4 Predictor Selection

It is common practice in regression analysis to fit various models that include subsets of the available predictor variables. However, for multi-category predictors like high school class, it is common practice to include a complete set of indicator variables so that each group is uniquely identified. Failure to keep a complete set of indicators is liable to misinterpretation, although it is perfectly acceptable if the interpretation is understood and appropriate for the data.

For example, consider again model (6), the ACOVA based on the factor high school class and the measurement variable age,  $x_5$ . Model (6) uses an intercept and indicators  $x_2, x_3, x_4$  for junior, sophomore, freshman, which makes senior the baseline category. Reasonable smaller models might be the simple linear regression

$$m(x) = \beta_1 + \beta_5 x_5,$$

which includes no effect on  $y$  due to high school class, and the one-way ANOVA model (4), which includes no effect for age. (Recall that high school class and age are closely related.)

It would be unusual to fit a model that does not include the intercept and all of  $x_2, x_3, x_4$ . For example, if we drop  $x_2$  from model (4) we get

$$m(x) = \beta_1 + \beta_3 x_3 + \beta_4 x_4 = \begin{cases} \beta_1, & \text{senior or junior} \\ \beta_1 + \beta_3, & \text{sophomore} \\ \beta_1 + \beta_4, & \text{freshman} \end{cases},$$

so we have created a model in which senior and junior have the same mean value. There is nothing wrong mathematically with such a model as long as it is understood and appropriate for the data.

An analysis of covariance without  $x_2$

$$m(x) = \beta_1 + \beta_3x_3 + \beta_4x_4 + \beta_5x_5$$

involves three parallel lines, one for senior and junior, one for sophomore, and one for freshman.

Alternatively, we might replace the intercept in model (4) and make  $x_1$  an indicator for senior. In other words, we consider model (5). If we drop  $x_2$  the model becomes

$$m(x) = \beta_1x_1 + \beta_3x_3 + \beta_4x_4 = \begin{cases} \beta_1, & \text{senior} \\ 0, & \text{junior} \\ \beta_3, & \text{sophomore} \\ \beta_4, & \text{freshman} \end{cases}.$$

This occurs because all of the predictor variables in the model take the value 0 for junior. If we incorporate the covariate age into this model we get

$$m(x) = \beta_1x_1 + \beta_3x_3 + \beta_4x_4 + \beta_5x_5 = \begin{cases} \beta_1 + \beta_5x_5, & \text{senior} \\ 0 + \beta_5x_5, & \text{junior} \\ \beta_3 + \beta_5x_5, & \text{sophomore} \\ \beta_4 + \beta_5x_5, & \text{freshman} \end{cases}$$

which are four parallel lines but sophomore has an intercept of 0.

Now suppose we wanted to construct a version of the ANOVA model (5) that treats freshmen and sophomore alike. We need a 0/1 indicator that identifies whether someone is either a freshman or a sophomore. Such an indicator is  $\tilde{x}_3 \equiv x_3 + x_4$ . Fitting the model

$$m(x) = \beta_1x_1 + \beta_2x_2 + \beta_3\tilde{x}_3 = \begin{cases} \beta_1, & \text{senior} \\ \beta_2, & \text{junior} \\ \beta_3, & \text{sophomore or freshman} \end{cases}$$

accomplishes our goal.

With  $x_1$  an indicator for senior, the ACOVA model (7) gives parallel lines for each group. We can also fit a model that gives completely separate lines for each group by multiplying the indicator variables by the continuous predictor age,  $x_5$ ,

$$\begin{aligned} m(x) &= \beta_{11}x_1 + \beta_{12}x_2 + \beta_{13}x_3 + \beta_{14}x_4 + \beta_{21}x_1x_5 + \beta_{22}x_2x_5 + \beta_{23}x_3x_5 + \beta_{24}x_4x_5 \\ &= \begin{cases} \beta_{11} + \beta_{21}x_5, & \text{senior} \\ \beta_{12} + \beta_{22}x_5, & \text{junior} \\ \beta_{13} + \beta_{23}x_5, & \text{sophomore} \\ \beta_{14} + \beta_{24}x_5, & \text{freshman} \end{cases}. \end{aligned}$$

When the categorical variable has quantitative values associated with its levels there are interesting relationships between regression and ANOVA models. Consider again our model

$$m(x) = \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 = \begin{cases} \beta_1, & \text{senior} \\ \beta_2, & \text{junior} \\ \beta_3, & \text{sophomore} \\ \beta_4, & \text{freshman} \end{cases}.$$

Associated with these high school classes are a number of years of schooling. Freshmen are in their ninth year of schooling while seniors are in their twelfth. We can define a “continuous” variable  $\tilde{x}$  that takes on the value of the number of years of education for each student. Clearly, if we treated  $\tilde{x}$  as a categorical variable, it would give the same results as high school class. Instead, we treat  $\tilde{x}$  as a continuous variable in which case the one-way ANOVA ends up being equivalent to the polynomial

$$m(\tilde{x}) = \gamma_1 + \gamma_2\tilde{x} + \gamma_3\tilde{x}^2 + \gamma_4\tilde{x}^3.$$

It is beyond the scope of this book to prove that these models are equivalent (see Christensen, 2002, Section 6.7) but they will give the same predictions for every individual in the data. We are using a third-degree polynomial because there are four distinct values of  $\tilde{x}$ . Two points determine a line, three points determine a parabola, and the four  $\tilde{x}$  values we have available determine a cubic polynomial that is equivalent to the one-way ANOVA. A primary difference in the models is that the ANOVA provides no way of making predictions for, say, 9.25 years of schooling while the cubic polynomial could easily do that. Unfortunately, just because it is easy to make such predictions does not mean that they are always sensible, see Christensen (1996, Section 7.11).

Finally, consider a regression model

$$m(x_i) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6}$$

and suppose that after inspecting the data, we decide to consider a model that involves constraints on the parameters:  $\beta_2 = 5$ ,  $\beta_4 = -\beta_3$ , and  $\beta_5 = 2\beta_6 + 3$ . Substituting into the original model gives

$$\begin{aligned} m(x_i) &= \beta_1 + 5x_{i2} + \beta_3 x_{i3} - \beta_3 x_{i4} + (2\beta_6 + 3)x_{i5} + \beta_6 x_{i6} \\ &= \beta_1 + 5x_{i2} + \beta_3(x_{i3} - x_{i4}) + \beta_6(2x_{i5} + x_{i6}) + 3x_{i5} \\ &= \gamma_1 + \gamma_2(x_{i3} - x_{i4}) + \gamma_3(2x_{i5} + x_{i6}) + (5x_{i2} + 3x_{i5}). \end{aligned}$$

This is just another regression model with different predictors that are functions of the original predictors. In the last equation the regression coefficients have been changed to  $\gamma$ s to emphasize that they are not the same coefficients as in the original model. The final regression term  $5x_{i2} + 3x_{i5}$  is a known constant that can be unique for each individual. Such a term is often called an *offset*.

#### 7.4.5 Several Categorical Covariates

Consider a linear regression model for a continuous response  $y$  with two categorical predictors: age (Young, Middle, Old) and sex. Together, these define six groups (Young, Male), (Middle, Male), (Old, Male), (Young, Female), (Middle, Female), and (Old, Female). With six groups we need six indicator functions. These are easily obtained from the indicator functions for the original two factors. Let  $x_1$  be the indicator for Young with  $x_2$  and  $x_3$  the indicators for Middle and Old, respectively. Similarly, let  $x_4$  indicate Male and  $x_5$  indicate Female. Taking products of the indicators from the categorical predictors, the indicator for (Young, Male) is  $\tilde{x}_1 \equiv x_1 x_4$  and the indicator for (Old, Female) is  $\tilde{x}_6 \equiv x_3 x_5$ . Fitting

$$m(x) = \beta_1 \tilde{x}_1 + \beta_2 \tilde{x}_2 + \beta_3 \tilde{x}_3 + \beta_4 \tilde{x}_4 + \beta_5 \tilde{x}_5 + \beta_6 \tilde{x}_6$$

or equivalently

$$m(x) = \beta_{11} x_1 x_4 + \beta_{12} x_1 x_5 + \beta_{21} x_2 x_4 + \beta_{22} x_2 x_5 + \beta_{31} x_3 x_4 + \beta_{32} x_3 x_5$$

involves fitting a one-way ANOVA on the six groups.

An alternative model with an interesting structure requires the effect of factors to be additive in the sense that whatever the age, the mean difference between Male and Female is the same, and whatever the sex, the differences among Young, Middle, and Old are the same. Thus, the effect of either factor does not depend on the other factor. *Interaction* is when the effect of one factor depends on the level of the other factor. To build a model with no interaction we incorporate an intercept but drop one indicator for each factor. Here we drop the indicators for Young and Male giving

$$m(x) = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_5. \tag{8}$$

The baseline group becomes (Young, Male). The relative effect for Middle is  $\beta_2$  so that  $\beta_2$  is added to the baseline for both Middle age groups. The relative effect for Old is  $\beta_3$  so that  $\beta_3$  is added to

Table 7.3: *Population means of the 6 groups.*

	Young	Middle	Old
Male	$\beta_1$	$\beta_1 + \beta_2$	$\beta_1 + \beta_3$
Female	$\beta_1 + \beta_4$	$\beta_1 + \beta_2 + \beta_4$	$\beta_1 + \beta_3 + \beta_4$

the baseline for both Old age groups. The relative effect for Female is  $\beta_4$ , so  $\beta_4$  is added to the baseline for all three Female groups. The results are summarized for the 6 groups in Table 7.3. We can compare groups by estimating mean differences and ratios.

#### EXERCISE 7.4.

- (a) Interpret the regression coefficients  $\beta_1, \beta_2, \beta_3, \beta_4$  in model (8).
- (b) What does the parameter  $\beta_2 - \beta_3 - \beta_4$  represent?
- (c) Is the parameter  $\beta_2 + \beta_3$  meaningful here? Explain.
- (d) Define a parameter that does not involve  $\beta_3$  that is not meaningful here.

EXERCISE 7.5. Consider a setting similar to Exercise 7.3 only now the predictor variables are sex and a three-category exposure variable (no exposure, low exposure, high exposure). The response variable is disease status ( $D = 1$  or  $D = 0$ ).

- (a) Write down the logistic regression model similar to the no interaction model (8) using (unexposed, female) as the baseline group.
- (b) Derive formulas for the probability of disease in each group.
- (c) Derive formulas for the odds of disease in each group.
- (d) Derive formulas for odds ratios and risk ratios comparing the 3 exposure groups to each other for males. Repeat for females. Comment.
- (e) Derive formulas for odds ratios comparing unexposed males to females, low exposed males to females, and high exposed males to females. Comment.

#### 7.4.6 Confounding

The basic idea of a predictive model is that if one thing happens it suggests that something else is more (or less) likely to happen. There is no suggestion that the first thing causes the second, only that they are likely to occur together.

One of the more dicey uses of regression is to evaluate the causal effect upon the response of some “treatment” variable. The logic is that if, after you properly adjust for everything else that could possibly cause the response, there still remains a relationship between the response and the treatment variable, then the treatment variable must be causing the relationship. For example, if you are interested in whether smoking causes lung cancer and after properly adjusting for every other variable that could possibly cause lung cancer, smoking still increases the predictive probability of having lung cancer, then smoking must cause lung cancer. Obviously the rub is in being sure that you have taken into account every other variable that could cause a susceptibility to lung cancer and even then, whether you have adjusted properly for their effects.

For example, if our FEV data were taken from a particular school district in which sports where very important, it might be the case that the school has a culture in which athletes are more likely to both smoke and take human growth hormone. In that case, even if we adjust for age, if we do not also adjust for taking human growth hormone, we may again see that smokers tend to have larger FEVs. In fact, larger FEVs might even occur without the human growth hormone based simply on the tendency of athletes to have higher FEVs. Our data contain no information about athletic activities and they certainly contain no information about illegally taking human growth hormone.

The problems involved in trying to adjust correctly for every variable correlated with the response are enormous.

Theoretically, it would be much simpler to run an experiment. To evaluate the effect of smoking on lung cancer, randomly assign people to smoking and nonsmoking, make sure they smoke the proper amount for twenty years, and then evaluate the incidence of lung cancer for the two groups. Randomly assigning the smoking treatment to people means that there should be no systematic differences between the two groups other than smoking. But, oh! Practically, legally, and morally we are not allowed to force people to smoke. We can perform similar experiments on laboratory animals and the smokers do get more cancer. (We leave to others the moral issue of forcing animals to smoke.)

Although trying to draw causal inferences from observational data is extremely difficult to do reliably, there is so much demand for it that great effort is put into doing it better. (We dare not say “doing it well.”) How often have you read news stories about some activity or food additive causing health results only to read later that the results have not been duplicated? Often these come from studies ascribing causation to predictive effects in an observational study.

A variable that is associated with both the response and the treatment variable, but is not caused by the treatment variable and vice versa, is called a *confounding variable*, or simply a *confounder*. For example, socioeconomic status is correlated with both lung cancer and smoking, and, at least in no obvious way does socioeconomic status cause smoking or vice versa. An apparent association between smoking and lung cancer may be enhanced, masked, or even reversed when socioeconomic status is ignored. The basic idea is simple: if a treatment variable  $x$  is correlated with a confounder  $c$ , and both  $x$  and  $c$  are correlated with the response  $y$ , then to isolate the effect of  $x$  on  $y$  we need to properly adjust for  $c$ .

The point of randomly assigning  $x$  to subjects in an experimental design is that there is then no basis for  $x$  to be correlated with any other variable (that is not determined by  $x$ ) and confounding variables should not exist. Whatever variables may be related to the behavior of subjects, they should be evenly distributed between the treatment groups when treatments are randomly assigned to subjects. This is precisely the philosophical basis on which one can infer causation in randomized experiments. When randomized, the treatment  $x$  is, unless one is unlucky with the randomization, the only thing that can cause the observed phenomena.

One method of adjusting for confounding variables is to match individuals—a case (treatment) with a control (non-treatment)—according to known confounders and then look at the difference in their responses. A simple illustration is to match smokers and nonsmokers of the same age and evaluate their lung cancer rates. However, residual confounding from unused, unmeasured, or unthought-of confounders may still be present. Alternatively, adjustments for confounders are made automatically when including the confounders along with the treatment variable in multi-predictor regression models. But again, the problem is whether we included all the necessary confounders and, even if we have, whether they are in a form that makes the appropriate adjustment for them.

Suppose we fit a model for a response variable (like FEV)

$$\tilde{m}(x) = \beta_1 + \beta_2 x_2 + \beta_3 x_3 \quad (9)$$

where  $x_2$  is a treatment variable (either continuous or binary like smoking) and  $x_3$  is another predictor, possibly a confounder (like age). Now consider two groups of subjects that have the same value for  $x_3$ , but their values for  $x_2$  differ by 1 unit. The difference in mean response between these two groups can be written

$$[\beta_1 + \beta_2(x_2 + 1) + \beta_3 x_3] - [\beta_1 + \beta_2 x_2 + \beta_3 x_3] = \beta_2.$$

The regression coefficient  $\beta_2$  is the difference in mean response associated with a 1 unit increase in  $x_2$ , when the other predictor in the model is held constant. In this sense we can isolate the effect of  $x_2$  because we are controlling for the other covariate. Using data to estimate the model parameters,

the estimated effect  $\hat{\beta}_2$  has been adjusted (or controlled) for  $x_3$ . This interpretation extends in simple ways to other models with multiple predictor variables.

**EXERCISE 7.6.** *Logistic Regression.* Carry out a similar explanation of how logistic regression odds ratios can be adjusted for confounders. Use an example (perhaps from your field of study) to aid in your description.

The problem is that, outside of experiments, we cannot look at the effect of changing  $x_2$  while holding  $x_3$  constant. In observational studies, we do not control the predictor variables, so within the data that generated the estimated regression coefficients, as  $x_2$  increases,  $x_3$  may naturally change also. The regression coefficients provide the combined effect of the two variables. Thus the data don't really give us a basis for discussing how changing  $x_2$  affects the response when other things are held constant. In the FEV example, the model describes the relationship between smoking, age, and FEV. It predicts what kind of FEV you can expect to have at your age if you happen to be a smoker or if you happen not to smoke. It does not address what happens to people who change from smokers to nonsmokers. To really find out what happens when you change, you have to look at data for people who change. But in the absence of data on people who change, this estimation of  $\beta_2$  may be the best information we can obtain.

An even bigger problem occurs when key variables have been left out of the model. Suppose the true regression model is

$$m(x) = \gamma_1 + \gamma_2 x_2 + \gamma_3 x_3 + \gamma_4 x_4.$$

Again, we imagine an FEV response with smoking and age as  $x_2$  and  $x_3$  and with  $x_4$  as either human growth hormone use or being an athlete. As discussed earlier for FEV, the treatment variable  $x_2$  may be correlated with  $x_4$ , which was not in our fitted model (9). The regression coefficient  $\beta_2$  implicitly incorporates effects from  $x_4$  as well as  $x_2$ , so changing  $x_2$  still may not give an accurate picture of the situation. This can even happen in a non-randomized “experiment” where we get to control  $x_2$  and  $x_3$ . Without randomization, we might subconsciously choose people for the smoking treatment who look more fit. We can only confidently imply causal effects from randomized experiments. Randomization gives us a basis for believing that the treatment variable  $x_2$  should be uncorrelated with any outside variable like  $x_4$ .

Despite these difficulties, predictive models that account for the treatment variable and all other predictor variables that could reasonably affect the response (and account for them appropriately) are typically the closest we can come to teasing out causation from these predictive models based on non-randomized data.

#### 7.4.7 Effect Modification/Interaction

A 2007 public health announcement stated that the risk of coronary heart disease is elevated for smokers, especially for those over age 35. This means that the increase in risk of coronary heart disease for smokers over nonsmokers is different depending on age. Younger smokers compared to younger nonsmokers are not as bad off as older smokers compared to older nonsmokers. Since the risk difference for smokers versus nonsmokers depends on age, we say that age *modifies the effect of smoking* or that age is an *effect modifier* of smoking. Medical researchers tend to use the term *effect modification*, while statisticians refer to *interaction* between the two risk factors. We use these terms interchangeably.

A predictor  $x_1$  is an effect modifier of a predictor  $x_2$  if the effect of  $x_2$  depends on the value of  $x_1$ . Note that  $x_1$  and  $x_2$  can both be continuous, both be categorical, or one of each. Figure 7.1 illustrates patterns of effect modification and the case of no interaction when  $x_1$  is continuous and  $x_2$  is binary, e.g., the FEV data. The key feature of interaction in this setting is that, as functions of  $x_1$ , the slopes of the regression functions are different depending on  $x_2$ . When there is no interaction the slopes are equal, i.e., the lines are parallel.

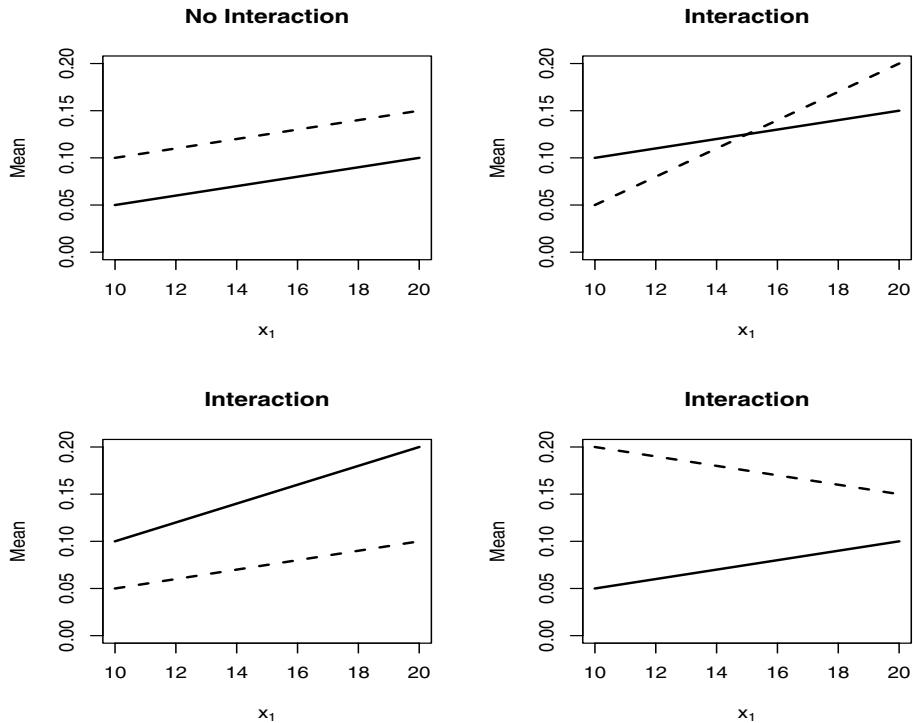


Figure 7.1: Patterns of effect modification between a continuous predictor  $x_1$  and a binary predictor  $x_2$ .

Modeling interaction in regression analysis often involves creating new predictors that are products of the interacting variables. New variables that are the product of two original variables are called two-way interaction terms, or simply interactions. Higher order interaction (e.g., interactions of 3 variables) can also be modeled but the resulting complex model may be difficult to interpret.

Three possible cases are described.

#### 7.4.7.1 Two Categorical Predictors

This is the simplest case. Consider a linear model for a continuous response  $y$  and two categorical predictors age (Young, Middle, Old) and sex. As discussed in Subsection 7.4.5, let  $x_1$ ,  $x_2$ , and  $x_3$  be indicators for Young, Middle, and Old, respectively, and let  $x_4$  indicate Male and  $x_5$  indicate Female. A model with no interaction is

$$m(x) = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_5.$$

As discussed earlier and illustrated in Table 7.3, the effect of an age category is the same regardless of sex and the effect of sex is the same regardless of age.

To allow interaction, we need a more general model. The most general model based on these two factors is one that allows each of the six groups (Young, Male), (Middle, Male), (Old, Male), (Young, Female), (Middle, Female), and (Old, Female) to have completely separate effects. As discussed earlier, such a model is achieved by

$$m(x) = \beta_{11}x_1x_4 + \beta_{12}x_1x_5 + \beta_{21}x_2x_4 + \beta_{22}x_2x_5 + \beta_{31}x_3x_4 + \beta_{32}x_3x_5.$$

**EXERCISE 7.7.** Using this  $m(x)$ , what parameter would you need to estimate if you wanted to quantify the difference in mean response between young males and young females? Old males and old females? Young males and old females?

**EXERCISE 7.8.** *Logistic Regression.* Carry out a similar explanation to Exercise 7.7 using odds ratios in logistic regression. Use an example (perhaps from your field of study) to aid in your description.

#### 7.4.7.2 One Continuous and One Categorical Predictor

First, what would it mean not to have interaction between a continuous and a categorical variable? It would mean that whatever relationship holds between the response variable and the continuous variable has the same form for every level of the categorical variable. For example, consider an ACOVA based on the factor high school class and the measurement variable age ( $x_5$ ). We use indicators  $x_1, x_2, x_3, x_4$  for senior, junior, sophomore, and freshman. A no interaction model that involves a linear effect for age is

$$m(x) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 = \begin{cases} \beta_1 + \beta_5 x_5, & \text{senior} \\ \beta_2 + \beta_5 x_5, & \text{junior} \\ \beta_3 + \beta_5 x_5, & \text{sophomore} \\ \beta_4 + \beta_5 x_5, & \text{freshman} \end{cases}.$$

This gives four parallel lines with the distances between the lines being the categorical effects. For a fixed age, the class effects are always the same. For a fixed class, the effect of age is always the same, increasing by a common slope  $\beta_5$ . More generally, we could write a no interaction model as

$$m(x) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + h(x_5) = \begin{cases} \beta_1 + h(x_5), & \text{senior} \\ \beta_2 + h(x_5), & \text{junior} \\ \beta_3 + h(x_5), & \text{sophomore} \\ \beta_4 + h(x_5), & \text{freshman} \end{cases}$$

where  $h$  is just some function of age. For example, it could be a cubic polynomial

$$\begin{aligned} m(x) &= \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_5^2 + \beta_7 x_5^3 \\ &= \begin{cases} \beta_1 + \beta_5 x_5 + \beta_6 x_5^2 + \beta_7 x_5^3, & \text{senior} \\ \beta_2 + \beta_5 x_5 + \beta_6 x_5^2 + \beta_7 x_5^3, & \text{junior} \\ \beta_3 + \beta_5 x_5 + \beta_6 x_5^2 + \beta_7 x_5^3, & \text{sophomore} \\ \beta_4 + \beta_5 x_5 + \beta_6 x_5^2 + \beta_7 x_5^3, & \text{freshman} \end{cases}. \end{aligned}$$

A natural way to model interaction is to use the same form for the relationship between  $y$  and  $x_5$ , but to allow the parameters to change depending on the group. For example, with a linear effect for  $x_5$  we might fit the model of completely separate lines for each high school class

$$\begin{aligned} m(x) &= \beta_{11} x_1 + \beta_{12} x_2 + \beta_{13} x_3 + \beta_{14} x_4 + \beta_{21} x_1 x_5 + \beta_{22} x_2 x_5 + \beta_{23} x_3 x_5 + \beta_{24} x_4 x_5 \\ &= \begin{cases} \beta_{11} + \beta_{21} x_5, & \text{senior} \\ \beta_{12} + \beta_{22} x_5, & \text{junior} \\ \beta_{13} + \beta_{23} x_5, & \text{sophomore} \\ \beta_{14} + \beta_{24} x_5, & \text{freshman} \end{cases}. \end{aligned}$$

For the cubic polynomial ACOVA model, an interaction model might take the form of completely separate cubics

$$\begin{aligned} m(x) &= \beta_{11} x_1 + \beta_{12} x_2 + \beta_{13} x_3 + \beta_{14} x_4 + \beta_{21} x_1 x_5 + \beta_{22} x_2 x_5 + \beta_{23} x_3 x_5 + \beta_{24} x_4 x_5 \\ &\quad + \sum_{r=1}^4 [\beta_{3r} x_r x_5^2 + \beta_{4r} x_r x_5^3] = \begin{cases} \beta_{11} + \beta_{21} x_5 + \beta_{31} x_5^2 + \beta_{41} x_5^3, & \text{senior} \\ \beta_{12} + \beta_{22} x_5 + \beta_{32} x_5^2 + \beta_{42} x_5^3, & \text{junior} \\ \beta_{13} + \beta_{23} x_5 + \beta_{33} x_5^2 + \beta_{43} x_5^3, & \text{sophomore} \\ \beta_{14} + \beta_{24} x_5 + \beta_{34} x_5^2 + \beta_{44} x_5^3, & \text{freshman} \end{cases}. \end{aligned}$$

But the most general interaction model allows completely different relationships between the response and the continuous predictor for each level of the factor

$$m(x) = \begin{cases} h_1(x_5), & \text{senior} \\ h_2(x_5), & \text{junior} \\ h_3(x_5), & \text{sophomore} \\ h_4(x_5), & \text{freshman} \end{cases}.$$

These models are limited only by the imagination of the data analyst.

**EXERCISE 7.9.** *Logistic Regression.* Carry out a similar explanation using odds ratios in logistic regression. Use an example (perhaps from your field of study) to aid in your description.

#### 7.4.7.3 Two Continuous Predictors

With two continuous predictors  $x_1$  and  $x_2$ , a no-interaction model takes the form

$$m(x) = h_1(x_1) + h_2(x_2).$$

In other words, the effect as  $x_1$  changes can be anything at all; it can be any function  $h_1$ , and similarly for  $x_2$ . However, the combined effect must be the sum of the two individual effects. The simplest common example of a no-interaction model is

$$m(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2. \quad (10)$$

Other common no-interaction models are a polynomial in  $x_1$  plus a polynomial in  $x_2$ .

An interaction model is literally any model that does not display the no-interaction structure. (The same could be said about our other two cases but with categorical variables we can identify some structure within the interaction models.) When generalizing no-interaction polynomial models, cross-product terms are often added to model interaction. For example, model (10) might be expanded to

$$m(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2. \quad (11)$$

A general additive, i.e., no interaction, polynomial model

$$m(x) = \beta_0 + \sum_{r=1}^R \beta_r x_1^r + \sum_{s=1}^S \beta_s x_2^s$$

might be extended to an interaction model

$$m(x) = \sum_{r=0}^R \sum_{s=0}^S \beta_{rs} x_1^r x_2^s.$$

The linear regression model (11) contains main effects for  $x_1$  and  $x_2$ , and the interaction term  $x_1 x_2$ . We have previously established that in model (10) without the interaction term,  $\beta_2$  denotes the difference in mean response for a 1 unit increase of  $x_2$ , controlled for  $x_1$ . Thus, a single parameter ( $\beta_2$ ) quantifies the effect of  $x_2$  adjusted for  $x_1$  in that model. What happens when we include the interaction term? From model (11), a 1 unit increase of  $x_2$  (controlled for  $x_1$ ) gives a change in mean response of

$$[\beta_0 + \beta_1 x_1 + \beta_2(x_2 + 1) + \beta_3 x_1(x_2 + 1)] - [\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2] = \beta_2 + \beta_3 x_1.$$

Here the effect of  $x_2$  depends on the value of  $x_1$ , as it should because this is exactly what occurs with effect modification. With an interaction term included in a regression model, there is no single parameter in the model that characterizes the main effect of  $x_2$ , and there shouldn't be because its

effect depends on the value of  $x_1$ . Statistical inferences would be derived from estimating  $\beta_2 + \beta_3 x_1$  across a range of values for  $x_1$  similar to those in the data.

**EXERCISE 7.10.** *Logistic Regression.* Carry out a similar explanation of how logistic regression odds ratios can model effect modification. Use an example (perhaps from your field of study) to aid in your description.

### 7.5 Illustration: FEV Data

We have used the FEV example to discuss general issues related to regression. We now look at what the actual data tell us.

We begin by ignoring the age variable. Write the difference in mean FEV for smokers and nonsmokers as  $\theta = \mu_S - \mu_{NS}$ . Using reference priors (see Exercise 7.11), the posterior median of  $\mu_S - \mu_{NS}$  is 0.15. Also,  $\Pr(\mu_S > \mu_{NS}|y) = 0.92$ , so we are 92% sure that smokers have higher mean FEV than nonsmokers. As mentioned before, this occurs because older people in the study are both more likely to smoke and more likely to have large FEV values. Indeed, we will see that age has a larger effect on FEV than smoking. Alternatively, we could have examined the relative FEV means,  $\mu_S/\mu_{NS}$ , or the percent difference in mean FEV,  $100(\mu_S - \mu_{NS})/\mu_S$  as parameters of interest. Estimating both of these parameters is easy in a Bayesian analysis.

**EXERCISE 7.11.** Obtain the two-sample FEV analysis using the following WinBUGS code. Modify the code to provide estimates of the other two parameters of interest mentioned.

```
model{
  for(i in 1:n){
    FEV[i] ~ dnorm(mu[Smoke[i]+1], tau[Smoke[i]+1])
  }
  mu[1] ~ dnorm(0,0.001)
  mu[2] ~ dnorm(0,0.001)
  tau[1] ~ dgamma(0.001,0.001)
  tau[2] ~ dgamma(0.001,0.001)
  theta <- mu[2]-mu[1] # smokers - nonsmokers
  thetadiff <- step(theta)
  junk1 <- Age[1]
}
```

*WinBUGS Tip.* You might wonder about the code that reads

```
junk1 <- Age[1]
```

WinBUGS requires all variables in the data to appear in the code, even variables that we don't want to use at the moment. To accommodate this we create a new variable named *junk1* that was assigned the value of the first person's age. We have used the variable *Age* in a way that does not alter the prior or likelihood so it has no impact on the results.

We now fit the ACOVA model that will provide an estimate of the effect of smoking after adjusting for age:

$$m(x) = \beta_1 + \beta_2 \text{Smoker} + \beta_3 \text{Age}.$$

Details of fitting such models are found in Chapter 9. Smokers have lower average FEV than nonsmokers,  $\Pr(\beta_2 < 0|y) = 0.95$ . The posterior median of  $\beta_2$  is  $-0.17$ . The age-adjusted mean FEV level for smokers is estimated as 0.17 liters lower than the mean FEV for nonsmokers. This is completely opposite from the unadjusted analysis.

The ACOVA models the relationship between FEV and age with parallel lines for smokers and nonsmokers. Because the lines are parallel, the effect of smoking in the model is simply the

distance between the two lines,  $\beta_2$ . To examine possible interaction, we now fit completely separate lines for smokers and nonsmokers. Since Smoker is a 0/1 variable, we can define a new 0/1 variable indicating Nonsmokers as  $1 - \text{Smoker}$ . One version of the two separate lines model is

$$\begin{aligned} m(x) &= \gamma_{11}\text{Nonsmoker} + \gamma_{12}\text{Smoker} + \gamma_{21}\text{Nonsmoker} \times \text{Age} + \gamma_{22}\text{Smoker} \times \text{Age} \\ &= \begin{cases} \gamma_{11} + \gamma_{21}\text{Age} & \text{Nonsmoker} \\ \gamma_{12} + \gamma_{22}\text{Age} & \text{Smoker} \end{cases}. \end{aligned}$$

Alternatively, we can fit the same model as

$$\begin{aligned} m(x) &= \beta_1 + \beta_2\text{Age} + \beta_3\text{Smoker} + \beta_4\text{Age} \times \text{Smoker} \\ &= \begin{cases} \beta_1 + \beta_2\text{Age}, & \text{Nonsmoker} \\ (\beta_1 + \beta_3) + (\beta_2 + \beta_4)\text{Age}, & \text{Smoker} \end{cases}. \end{aligned}$$

This alternative model has  $\beta_4$  as the difference in slope between smokers and nonsmokers. It turns out that  $\Pr(\beta_4 < 0|y) = 1$ , so that FEV increases at a slower rate for smokers than for nonsmokers. (Built into that statement was the fact, not in evidence, that  $\beta_2 + \beta_4 > 0$ .)

The Bayesian estimates of the regression lines presented in Figure 7.2 demonstrate that the effect of smoking is different for younger kids than for older kids. There appears to be no statistical difference between young smokers and young nonsmokers, but the smoking effect appears to become significant as kids age.

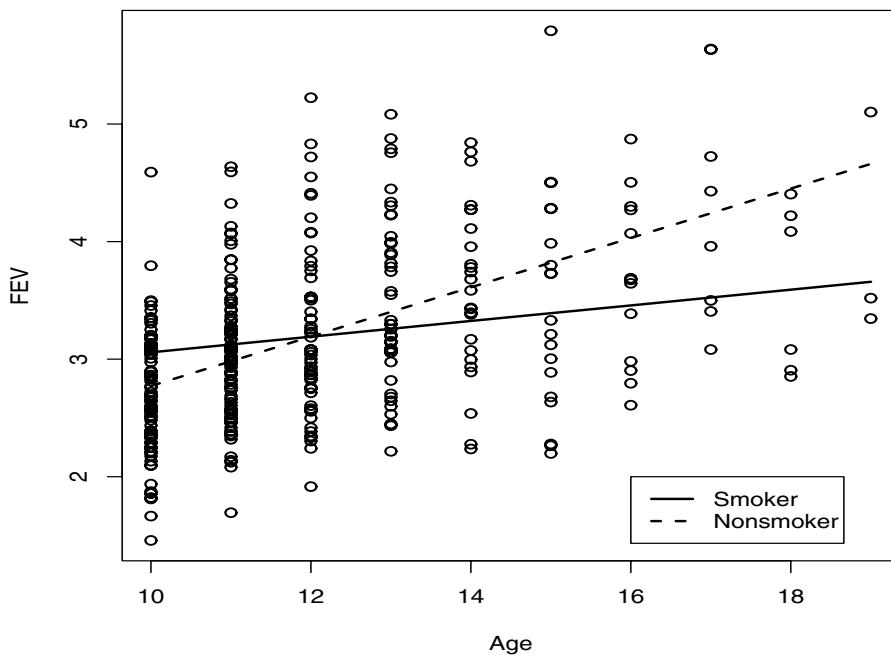


Figure 7.2: Interaction plot of the FEV data.

We should also note that there are relatively few older people in the data. That can influence the results. In addition, there is no particular reason to believe that FEV will increase linearly with age. Although we are adjusting for age, we may not be adjusting for age properly. We could fit polynomials in age or since age, although a measurement variable, is reported discretely, we could

Table 7.4: *Medians and 95% PIs for mean FEV of 11-year-old smokers and nonsmokers, and of 17 year old smokers and nonsmokers.*

Group	Population mean	Median (PI)
(11, Smoker)	$\beta_1 + \beta_3 + 11(\beta_2 + \beta_4)$	3.12 (2.87, 3.37)
(11, Nonsmoker)	$\beta_1 + 11\beta_2$	2.98 (2.90, 3.07)
(17, Smoker)	$\beta_1 + \beta_3 + 17(\beta_2 + \beta_4)$	3.52 (3.23, 3.82)
(17, Nonsmoker)	$\beta_1 + 17\beta_2$	4.24 (4.01, 4.47)

treat age as a multi-category factor. Also, there could be other predictors (e.g., height and sex) that are important. In Exercise 9.21, a detailed analysis of the full FEV data set is conducted.

Identifying functions of regression coefficients as parameters of scientific interest requires some care, especially in the presence of effect modification. Our central goal was to evaluate the association between smoking and FEV and since this analysis suggests that the association depends on age, there is effect modification. We need to estimate the smoking association for different age groups.

Proceeding with the separate lines model we have fitted, let's examine smoking effects for younger kids (age 11) and for older kids (age 17). The mean FEV for 11-year-old smokers is  $\beta_1 + \beta_3 + 11(\beta_2 + \beta_4)$  and for 11-year-old nonsmokers it is  $\beta_1 + 11\beta_2$ . Similar forms hold for 17-year-old kids. Posterior medians and 95% intervals for these parameters are easily obtained from WinBUGS and are reported for the 4 groups in Table .

The difference in mean FEV for 11-year-old smokers and nonsmokers is

$$[\beta_1 + \beta_3 + 11(\beta_2 + \beta_4)] - [\beta_1 + 11\beta_2] = \beta_3 + 11\beta_4.$$

This is the parameter from the model that represents the effect of smoking among 11 year olds. In general, the effect of smoking at age  $a$  is  $\beta_3 + a\beta_4$ . The posterior median for  $\beta_3 + 11\beta_4$  is 0.14 with 95% probability interval  $(-0.12, 0.41)$ . For  $\beta_3 + 17\beta_4$  we get  $-0.72$  and  $(-1.09, -0.34)$ . The evidence for a negative effect among smokers is substantial within the older age group but nonexistent within the younger age group. This could be the result of the cumulative effect of smoking; younger kids likely haven't been smoking as long as many older kids and smoking's effects may be cumulative.

Since the fitted model is based on data with  $10 \leq a \leq 19$ , we would not want to extrapolate too far beyond those ages. We have already called into question whether a linear effect for age is the appropriate way to adjust for age. Even if it is, a linear effect is highly unlikely to remain a reasonable approximation as age continues into the 20s, 30s, or 60s or for ages considerably younger than 10. These data were extracted from a larger set of data that contained children younger than 10, none of whom smoked. It would be surprising to see significant smoking effects at ages 9 or 10 since such children are unlikely to have been smoking long.

EXERCISE 7.12. Interpret the values of  $\tilde{\beta}_3 + 11\tilde{\beta}_4$  and  $\tilde{\beta}_3 + 17\tilde{\beta}_4$  in the FEV analysis where  $\tilde{\beta}_j$  denotes the posterior median of  $\beta_j$ .

---

## Chapter 8

---

# Binomial Regression

---

Binomial regression uses predictor variables to estimate the probability that some event will happen. Section 1 introduces three examples and discusses the sampling model. The first example examines the relationships between O-ring failures on space shuttles and launch temperatures. The second example looks at domestic and international students enrolled in graduate school. This is really a two independent binomials problem as discussed in Subsection 5.1.3 but here is discussed using the terminology of logistic regression. The final example involves survival of patients at a trauma center. The three examples are carried throughout the chapter. Section 2 presupposes that the priors are known and examines the posterior analysis. Section 3 considers two aspects of model checking: Box's significance test using the marginal distribution of the data and Bayes factors for examining alternative modeling strategies. Much of the heavy lifting occurs in Section 4, which discusses the choice of prior distributions. Section 5 examines random effects models for logistic regression. This chapter owes much to Bedrick, Christensen, and Johnson [BCJ] (1996, 1997) and to an unpublished early version of the 1997 article also discussed in Christensen (1997, Chapter 13).

### 8.1 The Sampling Model

We begin with two simple examples that involve only one predictor variable. Then we discuss the general model and an example with several predictors.

EXAMPLE 8.1.1. *O-Ring Data.* On January 28, 1986, the space shuttle *Challenger* crashed after takeoff. An explosion was caused by field O-rings that failed due to the low atmospheric temperature at takeoff, 31 degrees Fahrenheit. Table 8.1 gives data on O-ring failure  $y$  along with temperature at takeoff  $\tilde{x}$  from the 23 pre-*Challenger* space shuttle launches. Each flight is viewed as an independent trial. The result of a trial is 1 if any field O-rings failed on the flight and 0 if all the O-rings functioned properly. We want to estimate the probability of O-ring failure as a function of temperature. Previous analyses of these data include Dalal, Fowlkes, and Hoadley (1989), Lavine (1991), Martz and Zimmer (1992), and Christensen (1997).

Let  $\theta_i$  be the probability that any O-ring fails on launch  $i$ . We model the probability of O-ring failure as a simple function of temperature. As discussed in Chapter 7 we use a straight line to model a function of the probabilities

$$\text{logit}(\theta_i) \equiv \log[\theta_i/(1 - \theta_i)] = \beta_1 + \beta_2 \tilde{x}_i,$$

where  $\tilde{x}_i$  is the temperature corresponding to launch  $i$ . Defining

$$x_i = \begin{bmatrix} 1 \\ \tilde{x}_i \end{bmatrix} \quad \text{and} \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix},$$

write the model as  $\text{logit}(\theta_i) = x_i' \beta$ .

It is often convenient in regression problems to standardize continuous covariates so as to improve the performance of computational methods. The simplest standardization is to subtract the

Table 8.1: *O-ring failure data.*

Flight	Failure	Temperature	Flight	Failure	Temperature
14	1	53	17	0	70
9	1	57	2	1	70
23	1	58	11	1	70
10	1	63	6	0	72
1	0	66	7	0	73
5	0	67	16	0	75
13	0	67	21	1	75
15	0	67	19	0	76
4	0	68	22	0	76
3	0	69	12	0	78
8	0	70	20	0	79
			18	0	81

mean from any covariates. Writing  $\bar{x}_i = \sum_{i=1}^n \check{x}_i / n$ , the O-ring model becomes

$$y_i | \theta_i \stackrel{ind}{\sim} \text{Bin}(1, \theta_i), \quad \text{logit}(\theta_i) = \alpha + \beta_2(\check{x}_i - \bar{x}_i).$$

For the O-ring data, standardization is helpful but not necessary. In more complicated problems, numerical procedures may be difficult to perform without standardization. Standardization can also affect the specification of prior information.

Now consider an even simpler example. Its simplicity allows for easy illustration of relationships in logistic regression.

**EXAMPLE 8.1.2. Two Independent Samples.** We recruit statistics students into a graduate program from two populations: domestic students ( $i = 1$ ) and international students ( $i = 2$ ).  $N_1 = 10$  domestic students are accepted and  $N_2 = 10$  international students are accepted. We assume independence of all students, that  $\theta_1$  is the probability a domestic student enrolls in the program, and that  $\theta_2$  is the corresponding proportion among international students. We successfully recruit  $y_1 \sim \text{Bin}(N_1, \theta_1)$  domestic students and  $y_2 \sim \text{Bin}(N_2, \theta_2)$  international students.

Two independent samples is the simplest case of one-way analysis of variance (ANOVA). We can write a one-way ANOVA logit model

$$\text{logit}(\theta_i) \equiv \log[\theta_i / (1 - \theta_i)] = \mu + \alpha_i,$$

$i = 1, 2$ . This model has three parameters,  $\mu$ ,  $\alpha_1$ , and  $\alpha_2$ , to explain the behavior of two groups, so it is overparameterized. We re-parameterize as:

$$\text{logit}(\theta_1) = \beta_1, \quad \text{logit}(\theta_2) = \beta_1 + \beta_2.$$

Note that

$$\begin{aligned} \beta_2 &= \beta_1 + \beta_2 - \beta_1 \\ &= \text{logit}(\theta_2) - \text{logit}(\theta_1) \\ &= \log\left(\frac{\theta_2}{1 - \theta_2}\right) - \log\left(\frac{\theta_1}{1 - \theta_1}\right) \\ &= \log\left(\frac{\theta_2/[1 - \theta_2]}{\theta_1/[1 - \theta_1]}\right) \\ &= \log\left(\frac{\theta_2[1 - \theta_1]}{\theta_1[1 - \theta_2]}\right). \end{aligned}$$

Thus  $\beta_2$  is the difference in the log-odds between the two groups as well as the log of the odds ratio, i.e., the ratio of the odds of enrolling an international student relative to the odds of enrolling a domestic student.

This is also a simple linear regression model with a binary covariate. The model can be written

$$\text{logit}(\theta_i) = \beta_1 + \beta_2 \check{x}_i,$$

where the predictor variable  $\check{x}_i$  takes the value 0 if  $i = 1$  and 1 if  $i = 2$ . In vector notation the model is

$$\text{logit}(\theta_i) = x'_i \beta$$

where  $x'_1 = (1, 0)$ ,  $x'_2 = (1, 1)$ , and  $\beta = (\beta_1, \beta_2)'$ .

We now define the general binomial regression model. Consider data  $y_i$ ,  $i = 1, \dots, n$  with

$$y_i | \theta_i \stackrel{\text{ind}}{\sim} \text{Bin}(N_i, \theta_i).$$

Associated with each observation  $y_i$  are  $r$  fixed predictor variables  $x_{i1}, \dots, x_{ir}$ . Our models use linear functions of the predictor variables but, as discussed in Chapter 7, we need to take some function of the probabilities that transforms values between 0 and 1 into values on the real line. A convenient class of functions that does this consists of the inverses of cumulative distribution functions. If  $F$  is a cdf, its inverse is  $F^{-1}$ , that is  $F(F^{-1}(\theta)) = \theta$ . The general binomial regression model specifies

$$F^{-1}(\theta_i) = \sum_{j=1}^r \beta_j x_{ij} = x'_i \beta,$$

where  $x'_i = (x_{i1}, x_{i2}, \dots, x_{ir})$  and  $\beta = (\beta_1, \beta_2, \dots, \beta_r)'$ . Equivalently,

$$\theta_i = F(x'_i \beta).$$

Most often  $x_{i1} \equiv 1$  so that  $x'_i = (1, x_{i2}, \dots, x_{ir})$  and  $\beta_1$  is an intercept.

Three common selections for  $F$  correspond to *logistic*, *probit*, and *complementary log-log* regression models in which  $F$  is taken as one of

$$F(x' \beta) = \begin{cases} e^{x' \beta} / [1 + e^{x' \beta}] & \text{Logistic} \\ \Phi(x' \beta) & \text{Probit} \\ 1 - \exp[-e^{x' \beta}] & \text{Complementary log-log} \end{cases}.$$

Here  $\Phi(u)$  is the cdf of a standard normal distribution. One needs tables or a computer program to evaluate it. All three of these functions are strictly increasing and invertible. In this context,  $F^{-1}$  is often called a *link function*.

A complete analysis requires the use of both  $F$  and its inverse. The inverse of the logistic transform  $F(\eta) = e^\eta / [1 + e^\eta]$  is  $F^{-1}(\theta) = \log\{\theta/(1-\theta)\}$ , the *logit* transformation. For complementary log-log models,  $F^{-1}(\theta) = \log\{-\log(1-\theta)\}$ , hence its name. For probit models, one can only write  $F^{-1}(\theta) = \Phi^{-1}(\theta)$ . Since  $\Phi^{-1}$  cannot be written as a formula, it must be evaluated numerically, e.g., recall that  $\Phi^{-1}(0.5) = 0$  and  $\Phi^{-1}(0.95) = 1.645$ .

The likelihood function for independent binomial observations written as a vector  $y = (y_1, \dots, y_n)'$  is

$$\begin{aligned} L(\beta | y) \equiv \prod_{i=1}^n L(\beta | y_i) &= \prod_{i=1}^n \binom{N_i}{y_i} \theta_i^{y_i} (1 - \theta_i)^{N_i - y_i} \\ &= \prod_{i=1}^n \binom{N_i}{y_i} [F(x'_i \beta)]^{y_i} [1 - F(x'_i \beta)]^{N_i - y_i}. \end{aligned} \tag{1}$$

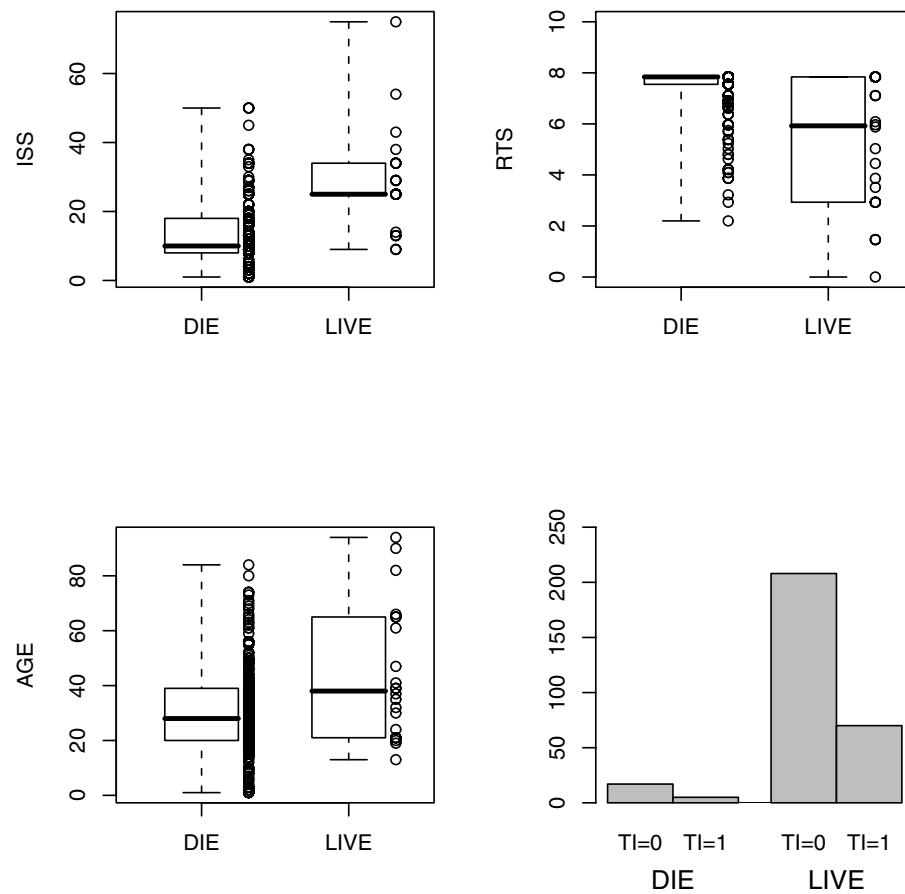


Figure 8.1: *Trauma data: Boxplots and a barchart.*

**EXAMPLE 8.1.3. *Trauma Data.*** Dr. Turner Osler, a trauma surgeon and former head of the Burn Unit at the University of New Mexico Trauma Center, provided data on survival of patients admitted to the University of New Mexico Trauma Center between 1991 and 1994. The predictor variables are injury severity score (ISS), revised trauma score (RTS), AGE, and type of injury (TI). TI is either blunt ( $TI = 0$ ), e.g., the result of a car crash, or penetrating ( $TI = 1$ ), e.g., gunshot or stab wounds. The ISS is an overall index of a patient's injuries based on the approximately 1,300 injuries catalogued in the Abbreviated Injury Scale. The ISS takes on values from 0 for a patient with no injuries to 75 for a patient with severe injuries in three or more body areas. The RTS is an index of physiologic status derived from the Glasgow Coma Scale. It is a weighted average of a number of measurements on an incoming patient such as systolic blood pressure and respiratory rate. The RTS takes on values from 0 for a patient with no vital signs to 7.84 for a patient with normal vital signs. We used a randomly selected subset of 300 observations. BCJ (1997) and Christensen (1997) provide similar analyses to that given here. In fact, only the computer programs have changed.

Figure 8.1 gives side-by-side boxplots comparing the 278 survivors and 22 fatalities on RTS, ISS, and AGE. Seventeen of the 225 patients with blunt injuries died. Five of the 75 patients with penetrating injuries died.

Dr. Osler proposed a logistic regression model to estimate the probability of a patient's death using an intercept; the predictors ISS, RTS, AGE, and TI; and an interaction between AGE and TI. AGE is viewed as a surrogate for a patient's physiologic reserve, i.e., their ability to withstand trauma. The model

$$\text{logit}(\theta_i) = \beta_1 + \beta_2 \text{ISS}_i + \beta_3 \text{RTS}_i + \beta_4 \text{AGE}_i + \beta_5 \text{TI}_i + \beta_6 (\text{AGE} * \text{TI})_i \quad (2)$$

is similar to logistic models used by trauma centers throughout the United States. (At least it was in 1996.) More generally, we could write the model

$$\text{logit}(\theta_i) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6},$$

where  $x_{i6} = x_{i4}x_{i5}$ . If we correct for the mean values of all the continuous predictor variables, the model becomes

$$\text{logit}(\theta_i) = \alpha + \beta_2(x_{i2} - \bar{x}_{.2}) + \beta_3(x_{i3} - \bar{x}_{.3}) + \beta_4(x_{i4} - \bar{x}_{.4}) + \beta_5 x_{i5} + \beta_6(x_{i4} - \bar{x}_{.4})x_{i5}.$$

With only the one binary variable TI, if we fix any two of the three continuous predictors in model (2), the log-odds as a function of the other variable will be one line for blunt injuries and a separate line for penetrating injuries. Because the model involves an  $\text{AGE} * \text{TI}$  interaction, as illustrated in Figure 8.2, with RTS fixed (here  $\text{RTS} = 3.34$ ) and ISS fixed ( $\text{ISS} = 40$ ), we get two completely separate lines as a function of AGE. For blunt injuries the line is

$$\text{logit}[\theta_B(\text{AGE})] = [\beta_1 + \beta_2 \times 40 + \beta_3 \times 3.34] + \beta_4 \text{AGE}$$

and for penetrating injuries the line is

$$\begin{aligned} \text{logit}[\theta_P(\text{AGE})] &= \text{logit}[\theta_B(\text{AGE})] + \beta_5 + \beta_6 \text{AGE} \\ &= [\beta_1 + \beta_2 \times 40 + \beta_3 \times 3.34] + \beta_4 \text{AGE} + \beta_5 + \beta_6 \text{AGE} \\ &= [\beta_1 + \beta_2 \times 40 + \beta_3 \times 3.34 + \beta_5] + (\beta_4 + \beta_6) \text{AGE}. \end{aligned}$$

Thus  $\beta_4$  is not really a regression coefficient for age, it is a regression coefficient for age *among* people having blunt injuries. The sum  $\beta_4 + \beta_6$  is the slope for the age line among people having penetrating injuries. Also,  $\beta_5$  is the difference between the intercepts for blunt and penetrating injuries, i.e., the difference (signed distance) between the lines when  $\text{AGE} = 0$ . Unless AGE has been standardized,  $\beta_5$  is not very interesting in itself. One would hope that these data, and therefore any fitted model, are not particularly relevant to people who are 0 years old. However, if  $\beta_6 = 0$  (no interaction between AGE and TI), the lines for blunt and penetrating injuries are parallel, in which case  $\beta_5$  is the difference between the lines regardless of AGE, so  $\beta_5$  is then interesting.

If we fix AGE and one other variable, and look at the lines as a function of the remaining variable, we get parallel lines with the distance between them depending not only on the type of injury but also on AGE. For example, if  $\text{RTS} = 3.34$  and  $\text{AGE} = 60$ , model (2) implies that for blunt injuries

$$\begin{aligned} \text{logit}[\theta_B(\text{ISS})] &= \beta_1 + \beta_2 \text{ISS} + \beta_3 \times 3.34 + \beta_4 \times 60 \\ &= [\beta_1 + \beta_3 \times 3.34 + \beta_4 \times 60] + \beta_2 \text{ISS} \end{aligned}$$

and for penetrating injuries

$$\begin{aligned} \text{logit}[\theta_P(\text{ISS})] &= \beta_1 + \beta_2 \text{ISS} + \beta_3 \times 3.34 + \beta_4 \times 60 + \beta_5 + \beta_6 \times 60 \\ &= [\beta_1 + \beta_3 \times 3.34 + \beta_5 + (\beta_4 + \beta_6) \times 60] + \beta_2 \text{ISS}. \end{aligned}$$

As a function of ISS, the slopes are the same, so the lines are parallel. However, the difference between the parallel lines is  $\beta_5 + \beta_6 \times 60$ , which obviously depends on AGE. Without an interaction between AGE and TI, the lines would remain parallel but the difference between them would only depend on TI. Such relationships are illustrated as part of the posterior analysis for the trauma data in Figure 8.7.

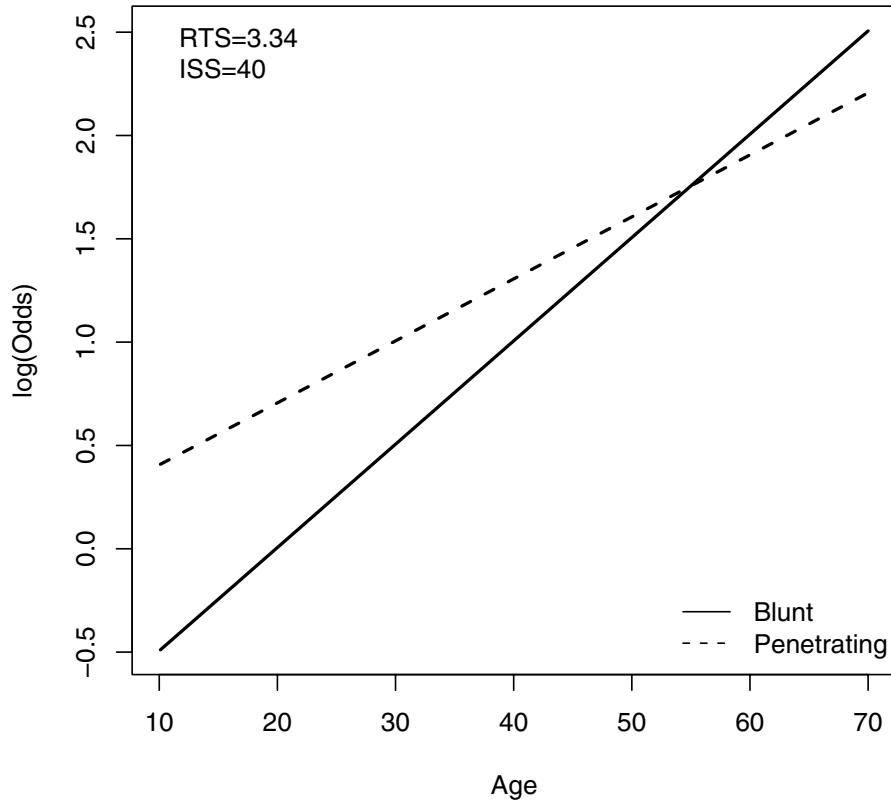


Figure 8.2: Trauma data: Logit model.

## 8.2 Binomial Regression Analysis

We now examine posterior analysis for our three examples. A Bayesian analysis requires specification of a prior distribution. For now, we simply present the analyses using subjective and reference priors. The rationale behind the priors is discussed in Section 4. Although it is convenient to write many formulae using a general cdf  $F$ , unless specifically stated otherwise, the examples use logistic models. Except in the first example, all posterior calculations are based on an MCMC sample from the posterior distribution of the regression coefficients, say,  $\{\beta^k : k = 1, \dots, m\}$ .

**EXAMPLE 8.2.1. Two Independent Samples.** This is by far the easiest of the three examples. Two independent samples were introduced in Example 3.1.3 and further explicated in Subsections 5.1.2 and 5.1.3.

For our student enrollment data, the sampling model is  $y_1|\theta_1 \sim \text{Bin}(N_1, \theta_1)$  for domestic students independent of  $y_2|\theta_2 \sim \text{Bin}(N_2, \theta_2)$  for international students. In Section 4 we choose

$$\theta_1 \sim \text{Beta}(11.26, 11.26) \perp\!\!\!\perp \theta_2 \sim \text{Beta}(13.32, 6.28).$$

The actual data are  $y_1 = 5, N_1 = 10, y_2 = 8, N_2 = 10$ . With independent data and independent priors, the posterior distributions are

$$\theta_1|y_1, y_2 \sim \theta_1|y_1 \sim \text{Beta}(16.26, 16.26) \perp\!\!\!\perp \theta_2|y_1, y_2 \sim \theta_2|y_2 \sim \text{Beta}(21.32, 8.28)$$

where, say,  $\theta_1|y_1 \sim \text{Beta}(11.26 + y_1, 11.26 + N_1 - y_1)$ . From Table 2.2, the posterior mean for  $\theta_1$  is 0.5 with a standard deviation of 0.086; the posterior mean of  $\theta_2$  is 0.72 with standard deviation 0.081.

The regression coefficients are a transformation of the  $\theta_i$ s,

$$\text{logit}(\theta_1) = \beta_1, \quad \text{logit}(\theta_2) = \beta_1 + \beta_2,$$

or

$$\beta_1 = \text{logit}(\theta_1), \quad \beta_2 = \text{logit}(\theta_2) - \text{logit}(\theta_1).$$

With a joint distribution for  $(\theta_1, \theta_2)'$  we could use Proposition B.4 to find the joint distribution of  $(\beta_1, \beta_2)'$ . Computer simulation provides a simpler approach. The posterior distribution of  $(\theta_1, \theta_2)'$  is a well-known distribution that is easy to sample. If we take a large random sample from the posterior, say,  $(\theta_1^k, \theta_2^k)', k = 1, \dots, m$ , we easily obtain a random sample from the joint posterior distribution of the regression coefficients by computing

$$\beta_1^k = \text{logit}(\theta_1^k), \quad \beta_2^k = \text{logit}(\theta_2^k) - \text{logit}(\theta_1^k).$$

For other functions, like the ratio of the probabilities  $\lambda = \theta_1/\theta_2$ , it is easy to sample from the posterior using  $\lambda^k = \theta_1^k/\theta_2^k$ . The proportion of these values that are less than 1 approximates the posterior probability that  $\lambda < 1$ , i.e., that domestic students are less likely to accept offers than international students.

**EXERCISE 8.2.1.** For Example 8.2.1, compare the exact posterior means and standard deviations of the  $\theta_j$ s with approximate posterior means and standard deviations obtained by simulation. Add the data to the WinBUGS code below.

```
model{
  for(i in 1:2){ y[i] ~ dbin(theta[i],N[i]) } # Likelihood
  theta[1] ~ dbeta(11.26,11.26) # The prior
  theta[2] ~ dbeta(13.32,6.28)
  beta[1] <- logit(theta[1]) # Induced posterior
  beta[2] <- logit(theta[2]) - logit(theta[1])
  lambda <- theta[1]/theta[2] # Induced posterior for ratio
  prob <- step(lambda -1) # Posterior prob. ratio > 1.
}
```

The odds ratio was defined and related to the regression coefficients in Section 1. Modify the WinBUGS code to obtain the posterior distribution of the odds ratio. Find the posterior mean and a 95% probability interval for the odds ratio.

**EXAMPLE 8.2.2. O-Ring Data.** Both the specification of the prior and the posterior computations are more complicated for the O-ring data. The prior is discussed in Section 4. The posterior computations are handled by the computer. The model is, for  $i = 1, \dots, 23$ ,

$$y_i|\theta_i \stackrel{\text{ind}}{\sim} \text{Bern}(\theta_i); \quad \text{logit}(\theta_i) = \beta_1 + \beta_2 \tilde{x}_i,$$

where  $\tilde{x}_i$  is the temperature at takeoff. Figure 8.3 gives contour plots of the prior and posterior distributions for  $\beta$ . The top plot is for this model using raw temperatures. The posterior creates the dark blob in the middle. The posterior shows much less variability. The prior and posterior both exhibit substantial correlation and appreciable skewness, with longer tails in the direction of small slopes and large intercepts. Most of the correlation is due to the parameterization.

Standardize the temperatures by subtracting  $\bar{x}_.$ , the sample mean of the temperatures, and dividing by the sample standard deviation  $s_x$ . Using the model

$$y_i|\theta_i \stackrel{\text{ind}}{\sim} \text{Bern}(\theta_i); \quad \text{logit}(\theta_i) = \beta_1 + \beta_2 (\tilde{x}_i - \bar{x}_.) / s_x,$$

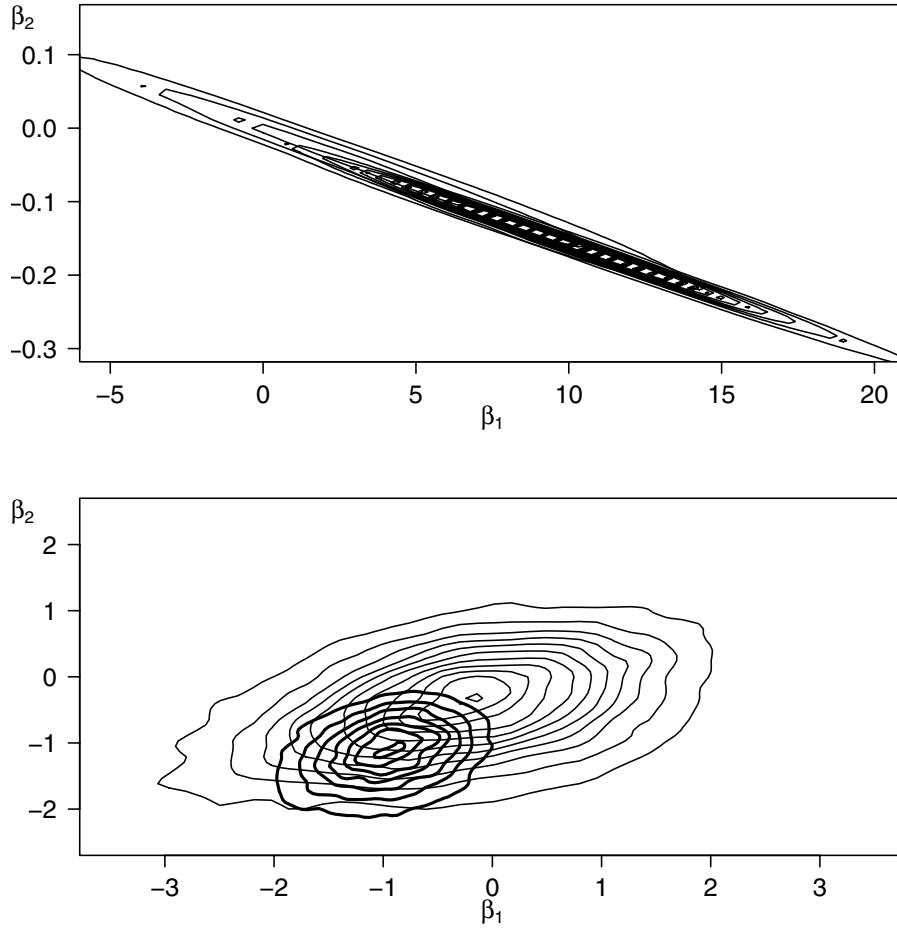


Figure 8.3 *O*-ring data: Prior and posterior contours for  $\beta$ . Top: raw temperatures. Bottom: standardized temperatures.

we see considerably less correlation between  $\beta_1$  and  $\beta_2$  in the bottom plot of Figure 8.3. The posterior consists of the contours towards the bottom left of the plot. The reduction in correlation is due to adjusting temperature for its mean. Subtracting  $\bar{x}_.$  does not change the meaning of  $\beta_2$ . Rescaling temperature merely rescales  $\beta_2$ . The standardization completely changes the meaning of  $\beta_1$ . Note the completely different scales for the two plots in Figure 8.3. Given the reduction in correlation observed here, it is not surprising that standardization frequently helps numerical properties when fitting models.

Subsequent subsections provide detailed analysis of the O-ring and trauma data. It would be traditional to begin with an examination of regression coefficients, but we think that predictive probabilities are more important.

### 8.2.1 Predictive Probabilities

The predictive probability of success in one new trial  $y_f$  with known covariate vector  $x \equiv x_f$  is

$$\Pr(y_f = 1|y) = E[F(x'\beta)|y] = \int F(x'\beta)p(\beta|y)d\beta. \quad (1)$$

The Monte Carlo approximation is

$$\Pr(y_f = 1|y) \doteq \frac{1}{m} \sum_{k=1}^m F(x'\beta^k).$$

Equation (1) has a direct interpretation as the predictive probability that the next trial will be a success, but it also has a secondary interpretation as the posterior mean of the parameter  $F(x'\beta)$ , which is the proportion of future trials that succeed. With the second interpretation, one may be interested in interval estimates. A 90% probability interval, say, is approximated by finding the appropriate sample percentiles of  $\{F(x'\beta^k) : k = 1, \dots, m\}$ .

**EXAMPLE 8.2.2 CONTINUED.** *O-Ring Data.* Figure 8.4 contains a plot of the predictive probabilities of O-ring failure along with 90% probability intervals. In other words, it gives  $E[F(x'\beta)|y]$  and a posterior interval estimate for  $F(x'\beta)$ . For low temperatures, the intervals are highly asymmetric because the posterior distribution of  $F(x'\beta)$  is highly skewed to the left. This also causes the mean  $E[F(x'\beta)|y]$  to be lower than the median.

The model indicates very high probabilities of O-ring failure at temperatures near the freezing point 32. The very large intervals for low temperatures occur because we are predicting well past the lowest temperature observation in the data, 53 degrees Fahrenheit.

**EXERCISE 8.2.** The WinBUGS code used to analyze the O-ring data follows. Independent Beta priors have been placed on the probabilities of O-ring failure at 55 and 75 degrees as discussed in Example 8.4.2. Run the WinBUGS code to obtain the predictive probabilities for the 23 launches. Modify the code to obtain (i) a point estimate and probability interval for the ratio and the difference between probabilities of O-ring failure at 55 and 75 degrees, respectively, (ii) the posterior probability that the probability of failure at 55 degrees is at least double the corresponding probability at 75 degrees, and (iii) the probability of O-ring failure at 31 degrees, the temperature on the day of the *Challenger* disaster. Remember that this is an extrapolation well beyond the range of temperatures in the data.

```

model{
  for(i in 1:n){
    y[i] ~ dbin(theta[i],1) # Likelihood
    logit(theta[i]) <- beta[1] + beta[2]*temp[i]
  }
  for(i in 1:2){ tildetheta[i] ~ dbeta(a[i],b[i]) } # Prior
  # Induced prior on the regression coefficients
  beta[1] <- (75/20)*logit(tildetheta[1])
  - (55/20)*logit(tildetheta[2])
  beta[2] <- (-1/20)*logit(tildetheta[1])
  + (1/20)*logit(tildetheta[2])
}
list(tildetheta=c(0.5,0.5))
# Sample size, hyperparameters of prior, and the data
list(n=23, a=c(1.6,1), b=c(1,1.6))
y[ ] temp[ ]
1 53
1 57
1 58
remaining 20 data lines go here
END

```

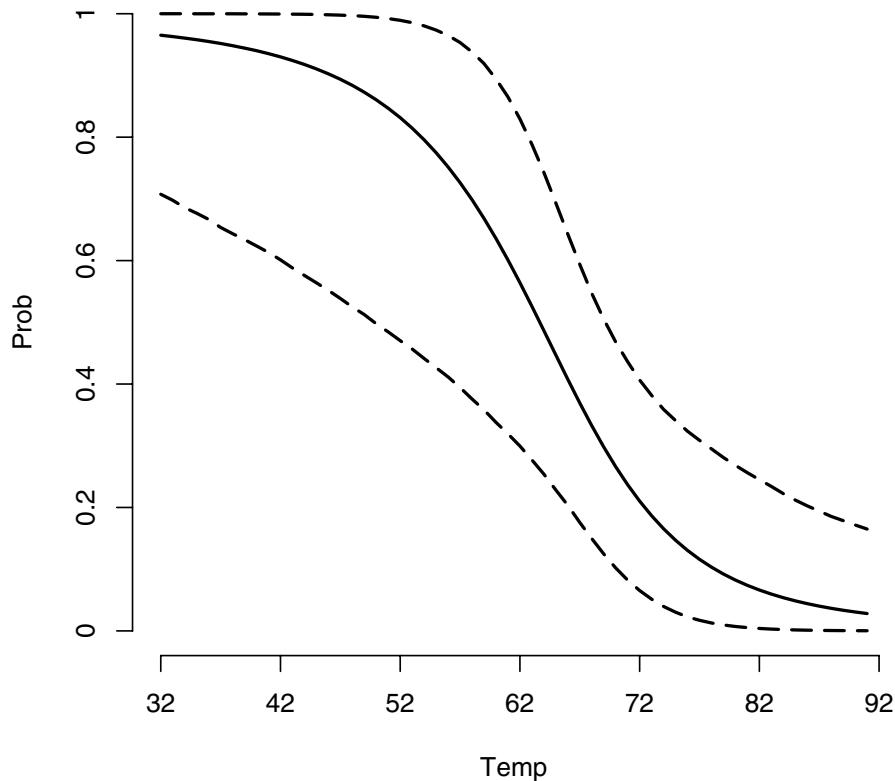


Figure 8.3: : *O-ring data: Predictive probabilities and 90% intervals*

**EXAMPLE 8.2.3. Trauma Data.** Using our informative prior from Section 4, Figure 8.5 presents predictive probabilities of death as a function of ISS for blunt and penetrating injuries. These are given for various values of RTS and AGE. Note that for 60-year-olds, there is essentially no difference in the probability of death due to blunt or penetrating injury. However, for 10-year-olds, the probability of death is higher for a penetrating injury. Thus, the effect of TI appears to depend on AGE. It is because we included a cross product (interaction) term for TI and AGE that this kind of “effect modification” can appear. If the interaction term were left out, the model would dictate that whatever TI effect might occur must be the same regardless of the value of AGE, and vice versa. Whether the interaction effect seen in Figure 8.5 is real depends on whether the corresponding regression coefficient  $\beta_6$  is different from zero. That issue is addressed in the next subsection.

### 8.2.2 Inference for Regression Coefficients

A component of  $\beta$ , say  $\beta_j$ , is typically estimated using either posterior medians or means. Approximating these from a posterior sample gives  $\tilde{\beta}_j \doteq \text{med}\{\beta_j^k\}$  and  $\hat{\beta}_j \doteq \sum_k \beta_j^k / m$ , respectively. Medians are preferred since they better represent the middle of skewed distributions. If the distribution is not skewed, they should be approximately the same. Posterior probability intervals are obtained using percentiles of the posterior distribution for each component. With a posterior sample of  $m = 10,000$ ,

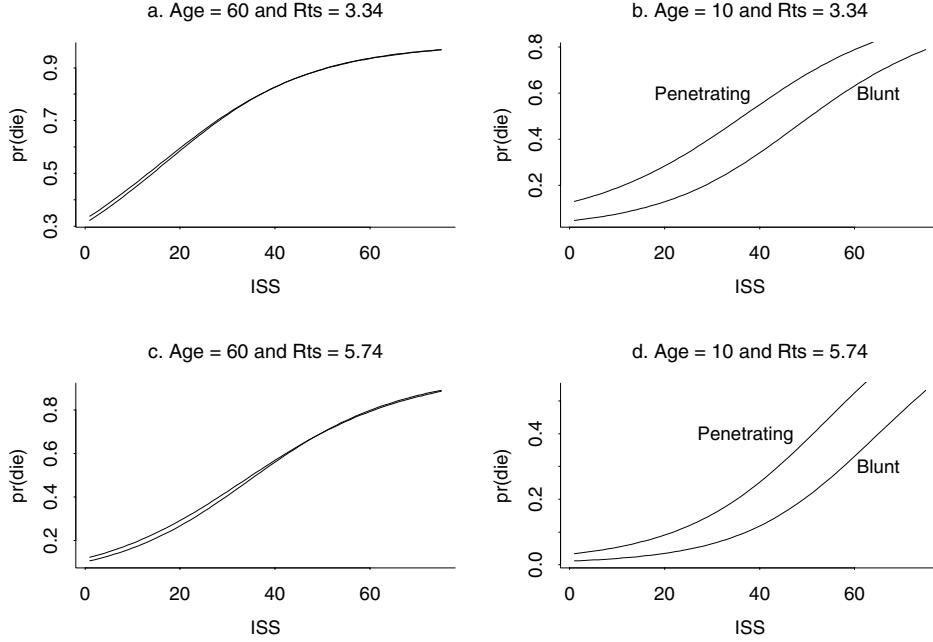


Figure 8.5: Trauma data: Predictive probabilities.

Table 8.2: O-ring data: Posterior summaries

Parameter	$\beta_1$	$\beta_2$
$\hat{\beta}_j = E(\beta_j y)$	10.86	-0.170
$sd(\beta_j y)$	4.70	0.069
5%	3.70	-0.292
25%	7.62	-0.213
$\tilde{\beta}_j = 50\%$	10.58	-0.166
75%	13.80	-0.123
95%	19.01	-0.065

order the  $\beta_j^k$ 's from smallest to largest to get ordered values  $\beta_j^{(k)}$ . A 95% probability interval for  $\beta_j$  is  $(\beta_j^{(250)}, \beta_j^{(9750)})$ . To make inferences for any function of  $\beta$ , say  $\gamma = g(\beta)$ , simply compute medians, means, and percentiles from the values  $\gamma^k = g(\beta^k)$ . In particular, one choice of  $g(\cdot)$  is  $g(\beta) = \beta_j$ .

**EXAMPLE 8.2.2 CONTINUED. O-Ring Data.** Table 8.2 presents posterior means, standard deviations, and percentiles of  $\beta_1$  and  $\beta_2$  for the O-ring data. The posterior means and medians are close, indicating some degree of symmetry. On the other hand, for  $\beta_1$ , the 75th and 95th percentiles are farther from the median than the 25th and 5th percentiles with the reverse being true for  $\beta_2$ . These inferences on regression coefficients are somewhat anticlimactic having already seen the dramatic effect of temperature in the predictive probability plot. Figure 8.6 gives the prior and posterior densities for  $\beta_2$  in the O-ring data. Consistent with Figure 8.3, the posterior density for  $\beta_2$  shows less variability and is centered on smaller values.

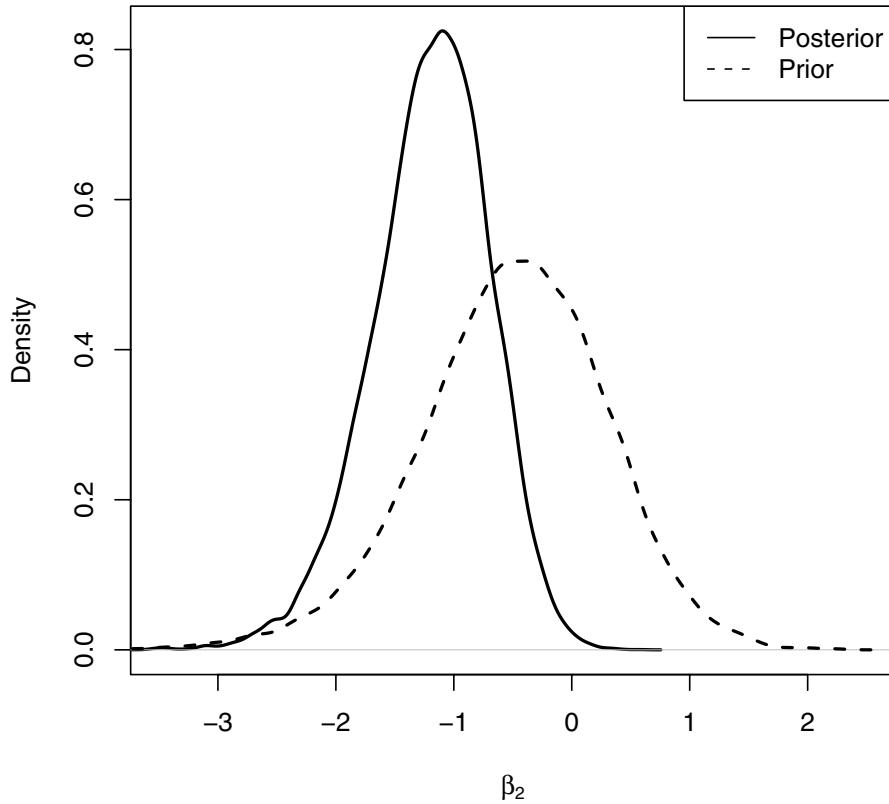


Figure 8.6: *O*-ring data: Prior and posterior densities for  $\beta_2$ .

In simple regression an unambiguous statement is frequently needed to the effect that  $\beta_2 \neq 0$ . Our posterior has  $\Pr(\beta_2 < 0|y) = 0.997$ . We are 99.7% sure that the slope is negative.

**EXERCISE 8.3.** Augment the code from Exercise 8.2 to calculate the predictive probability of at least one O-ring failure at 50 and 60 degrees, respectively, and to calculate the posterior probability that  $\beta_2 < -0.2$ . Obtain point and interval inferences for the proportions of shuttle flights with O-ring failures at 50 and 60 degrees, respectively. Make inferences for the corresponding odds ratio.

**EXAMPLE 8.2.3 CONTINUED.** *Trauma Data.* Table 8.3 presents posterior means, standard deviations, and percentiles from WinBUGS for the  $\beta_j$ 's of the trauma data model. These are based on an informative prior discussed in Section 4 and the improper flat prior  $p(\beta) = 1$ . For comparison, means and standard deviations obtained from importance sampling are also reported. The probability intervals for the  $\beta_j$ 's are about 3/4's as wide using the informative prior as with the flat prior. Low values of RTS are bad for the patient, so the tendency of the RTS coefficients to be negative is reasonable. In addition to the tabled results, the informative prior gives  $\Pr(\beta_2 > 0|y) > 0.99$ , suggesting that the coefficient of ISS is not zero.

In Figure 8.5 we plotted probabilities

$$\mathbb{E}[\theta(x)|y] \equiv \mathbb{E}[F(x'\beta)|y]$$

Table 8.3: *Trauma data: Posterior summaries*

Variable	Informative Prior						Imp. Samp.
	mean	sd	2.5%	median	97.5%	mean	
Intercept	-1.812	1.139	-4.049	-1.799	0.430	-1.79	1.10
ISS	0.065	0.021	0.026	0.065	0.107	0.07	0.02
RTS	-0.596	0.146	-0.895	-0.592	-0.324	-0.60	0.14
AGE	0.048	0.014	0.020	0.047	0.075	0.05	0.01
TI	1.171	1.06	-0.870	1.159	3.284	1.10	1.06
AGE × TI	-0.018	0.028	-0.073	-0.018	0.036	-0.02	0.03

Variable	Flat Prior						Imp. Samp.
	mean	sd	2.5%	median	97.5%	mean	
Intercept	-2.708	1.712	-6.036	-2.709	0.656	-2.81	1.60
ISS	0.085	0.029	0.029	0.085	0.142	0.09	0.03
RTS	-0.595	0.180	-0.962	-0.589	-0.26	-0.59	0.17
AGE	0.056	0.017	0.024	0.055	0.091	0.06	0.02
TI	1.409	1.42	-1.426	1.422	4.168	1.46	1.36
AGE × TI	-0.008	0.035	-0.081	-0.007	0.057	-0.01	0.03

for fixed values of AGE, RTS, both TI values, and a variety of ISS scores. Such plots give the most interpretable results. Figure 8.7 presents corresponding plots of the log-odds. While log-odds are less interpretable than probabilities, their geometry is simpler. All the curves are parallel lines with the distance between the lines depending on AGE.

In general, Figure 8.7 is plotting

$$E[F^{-1}\{\theta(x)\}|y] = E[x'\beta|y] = x'E[\beta|y] = x'\hat{\beta}.$$

Thus, the plots for the trauma data are obtained from the means in Table 8.3. In fact, Figure 8.2 was constructed in the same way as Figure 8.7 but with ISS fixed and AGE as the variable.

The AGE × TI coefficient could plausibly be 0, so there is reason to consider an alternative model in which for fixed ISS and RTS, the log-odds of death may be parallel lines in AGE, cf. the discussion of Figure 8.2. Although the probability intervals for TI also include 0, the weight of evidence is that the effect of TI is positive, causing the roughly parallel lines to be higher for  $TI = 1$ , penetrating injuries.

Finally, a word about sampling in Table 8.3. The results from importance sampling were based on a sample of size 10,000. The WinBUGS results were based on an MCMC sample of about 20,000. WinBUGS was quite slow fitting this sample because the predictor variables had not been standardized. We also computed the results for the informative prior based on an MCMC sample of about 80,000. This cut WinBUGS' reported Monte Carlo error by a factor of about 2. The output for the more extensive sample is

node	mean	sd	2.5%	median	97.5%
beta[1]	-1.757	1.145	-3.988	-1.756	0.488
beta[2]	0.065	0.021	0.024	0.065	0.106
beta[3]	-0.601	0.145	-0.897	-0.597	-0.326
beta[4]	0.047	0.014	0.021	0.047	0.075
beta[5]	1.099	1.089	-1.023	1.098	3.237
beta[6]	-0.017	0.028	-0.072	-0.016	0.037

Except for the intercept and TI (beta[5]), the 95% intervals are remarkably close to those in Table 8.3.

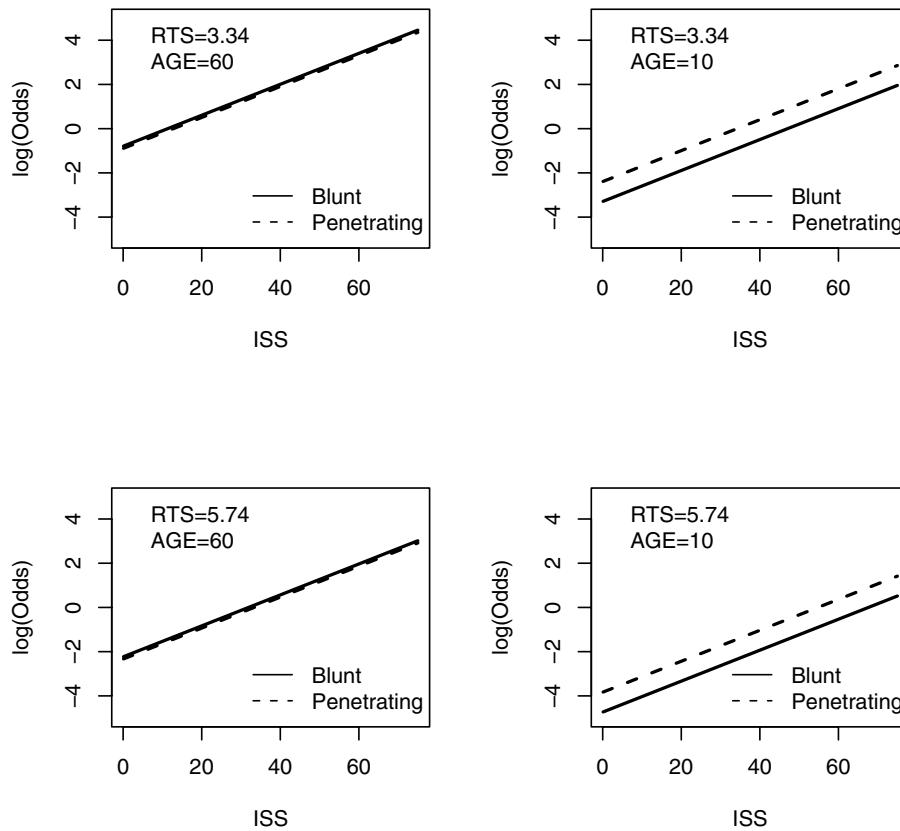


Figure 8.7: Trauma data: Posterior log-odds.

**EXERCISE 8.4.** The code given below models the probability of death while undergoing treatment for a traumatic injury using a flat prior on  $\beta$ . The data have the variable `death` listed with a 1 if the patient died and a 0 if they didn't. First, run the code and compare results with Table 8.3. Then modify the code to obtain the predictive probabilities of death for the covariate vectors  $(IS, RTS, AGE, TI) = (25, 5, 50, 1)$  and  $(50, 5, 50, 0)$ . Also obtain the posterior probabilities that each of the  $\beta_j$ 's is nonnegative. Obtain point and interval inferences for the proportions of deaths at the two covariate combinations above.

```

model{
  for(i in 1:n){
    death[i] ~ dbern(theta[i])
    logit(theta[i]) <- beta[1] + beta[2]*ISS[i] + beta[3]*RTS[i]
      + beta[4]*AGE[i] + beta[5]*TI[i]
      + beta[6]*AGE[i]*TI[i]
  }
  for(i in 1:6){ beta[i] ~ dflat() }
  junk <- ID[1]
}
list(beta= c(0,0,0,0,0,0))

```

Table 8.4: *O-ring data: Posterior summaries for  $LD_\alpha$ s.*

$\alpha$	Percentiles		
	5%	50%	95%
0.90	30.2	52.9	60.4
0.75	43.4	58.5	64.0
0.50	55.9	64.2	68.5
0.25	65.1	69.8	76.4
0.10	70.3	75.4	88.3

```

list(n=300)
ID[ ] death[ ] ISS[ ] TI[ ] RTS[ ] AGE[ ]
2979    0      1     1    7.8408   25
1167    0      9     0    7.8408   63
116     0     29     0    2.9304   32
remaining 297 data lines go here
END

```

### 8.2.3 Inference for $LD_\alpha$

With the O-ring data, it is of interest to estimate the temperature at which the chance of O-ring failure is, say 50%, or some other prespecified amount  $\alpha$ . This percentile is often called the  $LD_\alpha$  in bioassay problems. ( $LD$  denotes “lethal dose.”) It is obtained by solving  $\alpha = F(\beta_1 + x_\alpha \beta_2)$  for  $x_\alpha$ . The general solution is  $x_\alpha = [F^{-1}(\alpha) - \beta_1]/\beta_2$ . This is a simple function of  $\beta_1$  and  $\beta_2$ , so a posterior sample  $\{[F^{-1}(\alpha) - \beta_1^k]/\beta_2^k : k = 1, \dots, m\}$  is available. The posterior mean is unstable if  $\beta_2$  is near zero, so we recommend using the posterior median as a point estimate. In the special case of logistic regression,  $x_\alpha = [\text{logit}(\alpha) - \beta_1]/\beta_2$ .

Table 8.4 presents the posterior median and central 90% intervals for  $LD_\alpha$  using five values of  $\alpha$  for the O-ring data. In particular, the Bayesian analysis gives 69.8 degrees as the posterior median temperature at which the chance of O-ring failure is 0.25. The tails of the  $LD_\alpha$ s are very heavy due to a non-negligible probability of getting  $\beta_2$  values near zero.

**EXERCISE 8.5.** Modify the code in Exercise 8.2 to obtain inferences for the temperature at which 5% of flights would experience O-ring failures.

## 8.3 Model Checking

In this section, we examine Box’s (1980) method of model checking and the choice of an appropriate link function. In addition, BCJ (1997) and Christensen (1997, Subsection 13.3.1) discuss a variety of case deletion diagnostics for Bayesian binomial regression as a means of checking for outliers, see also Chaloner and Brant (1988). BCJ (1997) and Christensen (1997) also report results on sensitivity of the inferences to the choice of prior distributions.

### 8.3.1 Box’s Method

In Subsection 4.1.1 we examined a model-checking procedure due to Box (1980). This involved finding the probability that a new vector  $y$  has a marginal probability smaller than that of the vector  $y_{obs}$  that we actually observed, i.e., compute

$$\Pr[f(y) \leq f(y_{obs})],$$

where

$$f(y) = \int L(\beta|y)p(\beta)d\beta.$$

This gives a  $p$ -value, so that small values are of importance. For the O-ring data, this value is approximately 0.58. The probability is large, so there is no indication of a substantial problem with the model. If the improper flat prior  $p(\beta) = 1$  is used, the marginal distribution of the data  $f(y)$  may not exist.

Another model check considers Box's criterion one element of the  $y$  vector at a time, i.e., compute

$$\Pr [f(y_i) \leq f(y_{i,obs})].$$

This can be viewed as a Bayesian outlier check because we are assessing whether each observation is unusual relative to the model. For the O-ring data, all of these values are 1 except the two identical cases 13 and 14 that give 0.43 and case 18 that gives 0.37. This diagnostic gives no indication of problems with the model.

Rubin (1988) advocated Bayesian model checks using predictive rather than marginal distributions, see also Subsection 4.1.2. On the O-ring data, Rubin's analogues of the global and local model checks lead to identical conclusions. Similar methods also apply to the trauma data.

### 8.3.2 Link Selection

We now use Bayes factors to indicate which of the three link function models is most appropriate for the data: logistic ( $M_1$ ), probit ( $M_2$ ), or complementary log-log ( $M_3$ ). Bayes factors were introduced in Section 4.8. To compare models  $M_j$  and  $M_k$ , the Bayes factor is the number  $BF_{jk}$  that transforms the prior odds for the models into the posterior odds, i.e.,

$$\frac{\Pr(M_j|y)}{\Pr(M_k|y)} = [BF_{jk}] \frac{\Pr(M_j)}{\Pr(M_k)}.$$

In Section 4.8, we established that

$$BF_{jk} = \frac{f(y|M_j)}{f(y|M_k)}$$

where  $f(y|M_j)$  is the marginal density of obtaining  $y$  from model  $M_j$ . It is important to use informative priors when computing Bayes factors because otherwise, depending on the configuration of the data, the marginal distributions may not exist.

Our method for selecting informative prior distributions for regression problems discussed in the next section is particularly convenient when dealing with link selection problems. We elicit prior information on probabilities of success for various covariate values. This prior information is used to induce a prior distribution on the regression parameters. The proper interpretation of regression parameters depends on the link function that associates them with probabilities of success. With our method, one prior elicitation on success probabilities can be used to induce an appropriate and comparable prior distribution for the regression parameters of any link function. Computing  $f(y|M)$  for any model  $M$  involves integrating the corresponding likelihood function with respect to the induced prior on  $\beta$  for that model. Computational methods were given in Subsection 4.8.3. See Exercise 8.9 for inducing comparable priors for different link functions when computing Bayes factors for the O-ring data.

**EXAMPLE 8.2.3 CONTINUED.** *Trauma Data.* For the trauma data, the Bayes factors under our informative prior are  $BF_{21} = 1.05$ ,  $BF_{13} = 20.72$ , and, thus,

$$BF_{23} = BF_{21}/BF_{31} = BF_{21}BF_{13} = 1.05(20.72) = 21.83.$$

There is a suggestion against the complementary log-log model, but there is little to choose between the logistic and probit models.

## 8.4 Prior Distributions

In this section, we discuss informative, reference, and partially informative priors for binomial regression models. This involves placing a prior on the regression coefficients. Although regression coefficients provide a convenient way to specify a model, they do not provide the most intuitive way either to interpret the model or to specify prior information. We suggested earlier that model interpretations are best made by examining predictive probabilities rather than regression coefficients. Now we discuss methods of specifying prior information that focus on parameters that are more interpretable than regression coefficients and then using that prior information to induce a distribution on the regression coefficients.

### 8.4.1 Simple Regression

**EXAMPLE 8.4.1. Two Samples.** Consider again recruiting domestic ( $i = 1$ ) and international ( $i = 2$ ) graduate students. With  $y_i|\theta_i \sim \text{Bin}(N_i, \theta_i)$ , the model for the data is  $\text{logit}(\theta_i) = \beta_1 + \beta_2 \tilde{x}_i$  with  $\tilde{x}_i$  an indicator for international students. The graduate advisor has independent prior information for domestic students and international students. The advisor's best guesses for  $\theta_1$  and  $\theta_2$  are 0.5 and 0.7, respectively. Moreover, the advisor is 95% sure that  $\theta_1 \geq 0.33$  and 95% sure that  $\theta_2 \geq 0.5$ . We model the prior information using Beta distributions with the best guess taken to be the mode of the distribution. Using the methods of Section 5.1, we find that these characteristics correspond to

$$\theta_1 \sim \text{Beta}(11.26, 11.26) \quad \perp \quad \theta_2 \sim \text{Beta}(13.32, 6.28).$$

This joint distribution should be reconfirmed with the graduate advisor to verify that it adequately models their prior opinions.

The  $\theta_i$ s are parameters that are more directly related to potential observations than the regression coefficients, so it is easier to specify prior information for them. To obtain a prior for the regression coefficients, solve for the  $\beta_i$ s in terms of the  $\theta_i$ s:

$$\beta_1 = \text{logit}(\theta_1); \quad \beta_2 = \text{logit}(\theta_2) - \text{logit}(\theta_1).$$

The joint distribution on the  $\theta_i$ s determines a joint distribution on the  $\beta_i$ s through this relationship. Although knowledge about  $\theta_1$  is independent of knowledge of  $\theta_2$ , the  $\beta_i$ s are usually not independent.

Our scientific knowledge about the  $\theta_i$ s, which are easy to think about, directly determines a distribution for the  $\beta_i$ s, which are not so easy to think about. It is possible using Proposition B.4 to obtain the joint probability density for  $\beta$ , but since our approach is based on Monte Carlo simulations, that is unnecessary.

The independence assumption for the  $\theta_i$ s is a key part of the prior specification. With  $\theta_1$  and  $\theta_2$  independent, when told the value of  $\theta_1$ , we should not be inclined to revise our thinking about  $\theta_2$ . That seems reasonable if we are told that  $\theta_1$  is near its mode 0.5. However, it seems less reasonable if we are told, say, that  $\theta_1 \geq 0.95$ . Knowing that  $\theta_1 \geq 0.95$  would probably make us want to revise our distribution of  $\theta_2$  to make larger values more probable. However, 0.95 is 4.4 prior standard deviations above 0.5, so that  $\theta_1 \geq 0.95$  is extremely implausible under the current prior. The entire prior would need to be recalibrated if somehow it were now believed that  $\theta_1 \geq 0.95$  were true. If, after reflection, those situations that cause concern about independence are thought unlikely, we believe the independence assumption is reasonable.

Lack of independence can also occur if the international students were thought to be very similar to the domestic students regardless of the behavior of the domestic students. In that case, knowing  $\theta_1$  is informative about  $\theta_2$ , so our prior is not appropriate. A model that has  $\theta_2$  related to  $\theta_1$  is

$$F^{-1}(\theta_2) = F^{-1}(\theta_1) + \tilde{\varepsilon},$$

for a random error  $\tilde{\varepsilon}$  centered at 0. Another alternative is a hierarchical prior where  $\theta_i|\mu, \psi \stackrel{iid}{\sim} \text{Beta}(\mu \psi, (1-\mu) \psi)$ , with  $\mu \sim \text{Beta}(a, b)$  and perhaps a further distribution on  $\psi$ . This prior reflects

that the  $\theta_i$ s will be distinct but that if  $\psi$  is large, they will both be near  $\mu$ . The marginals for each of the  $\theta_i$ s are the same but the induced joint distribution reflects positive correlation, i.e., the  $\theta_i$ s are exchangeable.

The main idea in Example 8.4.1 was to specify prior distributions for  $\theta_1$  and  $\theta_2$  rather than on the regression parameters  $\beta_1$  and  $\beta_2$ . We do that because  $\theta_1$  and  $\theta_2$  have natural interpretations. In general a simple logistic regression has  $y|\theta \sim \text{Bin}(N, \theta)$  with

$$\text{logit}(\theta) = \beta_1 + \beta_2 \tilde{x}.$$

The intuitive parameter is  $\theta$ , the probability of success. The parameters  $\beta_1$  and  $\beta_2$  are a complex transformation of  $\theta$ . We put a prior distribution on  $\theta$  and let that distribution determine the prior on  $\beta_1$  and  $\beta_2$ . But  $\theta$  is actually a function of  $\tilde{x}$ ,  $\theta(\tilde{x})$ . There are two  $\beta_j$  parameters, so we need to specify a prior on two values of the function. Pick two values of  $\tilde{x}$ , say,  $\tilde{x}_1$  and  $\tilde{x}_2$ . Corresponding to these  $\tilde{x}$  values are two  $\theta$  values,  $\tilde{\theta}_1 \equiv \theta(\tilde{x}_1)$  and  $\tilde{\theta}_2 \equiv \theta(\tilde{x}_2)$ .  $\tilde{\theta}_h$  is the probability of success when the predictor variable is  $\tilde{x}_h$ , something that people can easily think about. In Example 8.4.1, it was natural to let  $\tilde{x}_1 = 0$  and  $\tilde{x}_2 = 1$ , so that  $\tilde{\theta}_h = \theta_h$ , but if  $\tilde{x}$  is continuous, more thought is required in choosing the  $\tilde{x}_h$ s.

**EXERCISE 8.6.** WinBUGS code for inducing the prior on  $\beta$  in Example 8.4.1 is:

```
model{
  theta[1] ~ dbeta(11.26,11.26)
  theta[2] ~ dbeta(13.32,6.28)
  beta[1] <- logit(theta[1])
  beta[2] <- logit(theta[2]) - logit(theta[1])
}
```

Monitor both  $\theta$  and  $\beta$  and visualize the prior on the  $\theta_i$ s and the induced prior on the  $\beta_j$ s, cf. Figure 8.3. Repeat this process using  $\theta_1 \sim \text{Beta}(6,6)$  and  $\theta_2 \sim \text{Beta}(6.5,3.25)$ . Also repeat it with  $\theta_1 \sim \text{Beta}(22,22)$  and  $\theta_2 \sim \text{Beta}(26,13)$ . Describe how the priors and induced priors differ.

**EXAMPLE 8.4.2.** *O-Ring Data.* Consider fitting a simple binomial regression model on temperature to the data in Table 8.1,  $F^{-1}(\theta_i) = \beta_1 + \beta_2 \tilde{x}_i = x'_i \beta$ . Our prior is defined by giving independent distributions to the probabilities of O-ring failure at temperatures  $\tilde{x}_1 = 55$  and  $\tilde{x}_2 = 75$  degrees Fahrenheit. In a slight abuse of notation, write the vector  $(1, \tilde{x}_h)$  as  $\tilde{x}'_h$ , so

$$\beta_1 + \beta_2 \tilde{\theta}_h = (1, \tilde{x}_h) \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \tilde{x}'_h \beta$$

and define  $\tilde{\theta}_1$  and  $\tilde{\theta}_2$  by  $\tilde{\theta}_h \equiv \theta(\tilde{x}_h) = F(\tilde{x}'_h \beta)$ . The  $\tilde{x}_h$ s should be chosen within the range of the observed temperatures but far enough apart so that information about the corresponding probabilities can be reasonably assumed independent. The selected temperatures should also be amenable to expert opinion.

For the O-ring data, the priors on  $\tilde{\theta}_1 = \theta(55)$  and  $\tilde{\theta}_2 = \theta(75)$  were chosen to have modes of one and zero, respectively, and to have  $\Pr(\tilde{\theta}_1 > 1/2) = 2/3$  and  $\Pr(\tilde{\theta}_2 < 1/2) = 2/3$ . Our priors on  $\tilde{\theta}_1$  and  $\tilde{\theta}_2$  are Beta(1.6, 1) and Beta(1, 1.6), respectively. This reflects a mildly informative belief that the probability of failure is higher at low temperatures than at high temperatures, and will result in an induced prior on  $\beta_2$  such that  $\Pr(\beta_2 < 0) \doteq 0.74$ .

BCJ, (1997) as reported in Christensen (1997, Chapter 13), used similar but distinct priors,  $\tilde{\theta}_1 \sim \text{Beta}(1, 0.577)$  and  $\tilde{\theta}_2 \sim \text{Beta}(0.577, 1)$ . These priors also have  $\Pr(\tilde{\theta}_1 > 1/2) = 2/3$  and  $\Pr(\tilde{\theta}_2 < 1/2) = 2/3$  and the modes are also at one and zero. The BCJ priors have a “J” shape, i.e., are convex. The priors used here are concave. In our priors, the values of the prior density at the modes are finite. In the BCJ priors, the densities at the modes are infinite.

The induced prior on  $\beta = (\beta_1, \beta_2)'$  can be, in principle, determined using Proposition B.4, but we approximate it using the posterior sample. Under the logistic model, this prior on  $\beta$  can be shown to be a data augmentation prior (DAP) in the sense of Section 2.3 in that it has the same functional form as the likelihood

$$p(\beta) \propto \prod_{i=1}^2 [F(\tilde{x}'_i \beta)]^{\tilde{y}_i} [1 - F(\tilde{x}'_i \beta)]^{\tilde{N}_i - \tilde{y}_i}, \quad (1)$$

where  $\tilde{N}_1 = \tilde{N}_2 = 2.6$ ,  $\tilde{y}_1 = 1.6$ , and  $\tilde{y}_2 = 1$ . (See Exercise 8.8 for the derivation.) With this DAP, the prior on  $\tilde{\theta}_1$  can be thought of as 1.6 prior O-ring failures out of 2.6 trials at  $\tilde{x}_1 = 55$ , and for  $\tilde{\theta}_2$  it can be thought of as 1 prior O-ring failure out of 2.6 trials at  $\tilde{x}_2 = 75$ . The weight attached to the prior is equivalent to  $\tilde{N}_1 + \tilde{N}_2 = 5.2$  “prior” observations. The posterior density for  $\beta$  also has the same functional form as the likelihood, i.e.,

$$p(\beta|y) \propto \prod_{i=1}^n [F(x'_i \beta)]^{y_i} [1 - F(x'_i \beta)]^{N_i - y_i} \prod_{i=1}^2 [F(\tilde{x}'_i \beta)]^{\tilde{y}_i} [1 - F(\tilde{x}'_i \beta)]^{\tilde{N}_i - \tilde{y}_i}.$$

While of some interest mathematically, this preservation of the functional form has little bearing on our computational methods.

#### 8.4.2 General Regression

We begin with a discussion of prior elicitation for general models and then examine the trauma data. Historically, the prior for  $\beta$  has been chosen as a multivariate normal distribution or as some reference distribution such as the improper flat prior  $p(\beta) = 1$ . These are convenient in large sample situations where both the likelihood and the posterior for  $\beta$  are approximately normal. Instead, we focus on the assessment of “success” probabilities for various choices of covariate values.

The genesis of this approach lies with Tsutakawa (1975), Tsutakawa and Lin (1986), and Grieve (1988) who considered independent prior distributions on two probabilities of “success” in simple linear binomial regression problems. Tsutakawa and Lin (1986) argued that eliciting information about success probabilities should be much easier than eliciting information about regression coefficients. Obviously, we heartily agree. This seems clearly true if one entertains the possibility of two or more models, such as logistic regression versus probit or complementary log-log regression. The regression coefficients for these three models require separate elicitations, whereas if one has elicited a prior for probabilities, it is straightforward to induce the requisite prior on  $\beta$  for each model (see also Subsection 8.3.2). BCJ (1996, 1997) give further details on this approach to specifying priors for regression problems, including discussions of priors with order restrictions on the  $\tilde{\theta}$ s.

In binomial regression with  $r$  predictor variables, we have independent observations  $y_i|\theta_i \sim \text{Bin}(N_i, \theta_i)$  and  $\theta_i = F(x'_i \beta)$ . We need a prior on  $\beta$ . Rather than directly eliciting an  $r$  dimensional prior on  $\beta$ , we pick  $r$  predictor variable vectors  $\{\tilde{x}_h : h = 1, \dots, r\}$ , imagine the associated probabilities  $\tilde{\theta}_h = F(\tilde{x}'_h \beta)$ , and elicit independent prior distributions on the  $\tilde{\theta}_h$ s. We typically use

$$\tilde{\theta}_h \sim \text{Beta}(\tilde{y}_h, \tilde{N}_h - \tilde{y}_h)$$

regardless of the choice of the link function  $F^{-1}$ . While we regard information about the  $\tilde{\theta}_h$ s independently, the approach can, in theory, be carried out with any joint distribution for the  $\tilde{\theta}_h$ s. The problem is less in doing the calculus, than in specifying a realistic joint distribution when the independence assumption is not appropriate. BCJ (1996) discuss the justification for the independence assumption. For moderate to large  $r$ , the process of prior elicitation can become quite tedious, so later we discuss the use of partial prior information in which we specify prior information at fewer than  $r$  vectors  $\tilde{x}_h$ .

Selection of the  $\tilde{x}_h$ s is key. Choosing the  $\tilde{x}_h$ s is simply a question of experimental design. We are asking the expert to perform a thought experiment, so statistical experimental design can be helpful

in this process. There are numerous books available on experimental design. One of our favorites, or rather, the favorite of one of us, is Christensen (1996).

Define the matrix  $\tilde{X}' = (\tilde{x}_1, \dots, \tilde{x}_r)$ . This must be nonsingular so that we can find its inverse. In particular, with our regression parameter vector  $\beta$  and defining the vector  $\tilde{\theta} = (\tilde{\theta}_1, \dots, \tilde{\theta}_r)'$ , our model specifies that

$$\tilde{\theta} \equiv \begin{bmatrix} \tilde{\theta}_1 \\ \vdots \\ \tilde{\theta}_r \end{bmatrix} = \begin{bmatrix} F(\tilde{x}'_1 \beta) \\ \vdots \\ F(\tilde{x}'_r \beta) \end{bmatrix} = F(\tilde{X} \beta).$$

Here it is understood that when we apply a function originally defined on real numbers to a vector, we apply it to each element of the vector. Solving this equation for  $\beta$  gives

$$\beta = \tilde{X}^{-1} F^{-1}(\tilde{\theta}). \quad (2)$$

Thus, the distribution that we elicit on  $\tilde{\theta}$  determines a distribution on  $\beta$ . BCJ (1996) recommend calculating the condition number of  $\tilde{X}$  to ascertain that the chosen  $\tilde{x}_i$ s are neither too close nor too far apart. Belsley (1991) discusses condition numbers.

In any case, as indicated above, the distribution on the  $\tilde{\theta}_h$ s induces a distribution on  $\beta$  through the standard change of variable formula Proposition B.4. Fortunately, when simulating the posterior, we do not need to actually perform the change of variable calculus, we merely sample from the distribution of  $\tilde{\theta}$  and use (2) to obtain a sample of  $\beta$ s. Nonetheless, the prior distribution has the functional form

$$p(\beta) \propto \prod_{h=1}^r [F(\tilde{x}'_h \beta)]^{\tilde{y}_h - 1} [1 - F(\tilde{x}'_h \beta)]^{\tilde{N}_h - \tilde{y}_h - 1} f(\tilde{x}'_h \beta), \quad (3)$$

where  $f(\cdot)$  is the first derivative of  $F(\cdot)$ . In the case of logistic regression, the prior simplifies to

$$p(\beta) \propto \prod_{h=1}^r [F(\tilde{x}'_h \beta)]^{\tilde{y}_h} [1 - F(\tilde{x}'_h \beta)]^{\tilde{N}_h - \tilde{y}_h}, \quad F(u) = \frac{e^u}{1 + e^u}, \quad (4)$$

which has the same form as the likelihood function, so we have a DAP as mentioned in Section 2.3.

**EXERCISE 8.7.** *O-Ring Data.* Write down the explicit form of equation (2) for Example 8.4.2. Show that the solution is

$$\beta = \begin{pmatrix} (75/20) \text{logit}(\tilde{\theta}_1) - (55/20) \text{logit}(\tilde{\theta}_2) \\ (-1/20) \text{logit}(\tilde{\theta}_1) + (1/20) \text{logit}(\tilde{\theta}_2) \end{pmatrix} = \begin{pmatrix} (75/20) & -(55/20) \\ (-1/20) & (1/20) \end{pmatrix} \text{logit}(\tilde{\theta}),$$

using (i) simple algebra and (ii) matrix algebra.

**EXAMPLE 8.4.3.** *Trauma Data.* To induce an informative prior distribution on the  $r = 6$  dimensional vector  $\beta$ , we require a joint distribution on death probabilities for 6 sets of conditions  $\tilde{x}'_h = (1, ISS_h, RTS_h, AGE_h, TI_h, AGE_h \times TI_h)$ . The  $\tilde{x}_h$ s along with the parameters of the Beta priors for the  $\tilde{\theta}_i$ s are given in Table 8.5.

With four distinct predictor variables, our expert Dr. Osler selected four “comfortable” covariate combinations. The idea was to pick values of the variables that were relatively extreme within the data but still had substantial probabilities for both success and failure. Dr. Osler had relatively little difficulty determining priors for these first four combinations. However, since our fitted model also included an intercept and an interaction, we needed  $r = 6$  covariate combinations for eliciting priors. We noted that the first four combinations constituted a 1/4th rep. of a  $2 \times 2 \times 2 \times 2$  factorial design having ISS at levels 25 and 41, RTS at levels 3.34 and 7.84, AGE at levels 10 and 60, and TI at levels 0 and 1. To this we added two “center” points. The “center” points differ on the TI scale since an average TI value is nonsense. This resulted in a nonsingular  $\tilde{X}$  matrix as required.

Table 8.5: Trauma data: Prior specification.

$h$	Design for Prior						$\tilde{y}_h$	$\tilde{N}_h - \tilde{y}_h$	Prior Median
	$\tilde{x}'_h$	$\tilde{N}_h$	$\tilde{y}_h$	$\tilde{N}_h - \tilde{y}_h$	$\tilde{y}_h$	$\tilde{N}_h - \tilde{y}_h$			
1	1	25	7.84	60	0	0	1.1	8.5	0.09
2	1	25	3.34	10	0	0	3.0	11.0	0.20
3	1	41	3.34	60	1	60	5.9	1.7	0.80
4	1	41	7.84	10	1	10	1.3	12.0	0.07
5	1	33	5.74	35	0	0	1.1	4.9	0.15
6	1	33	5.74	35	1	35	1.5	5.5	0.19

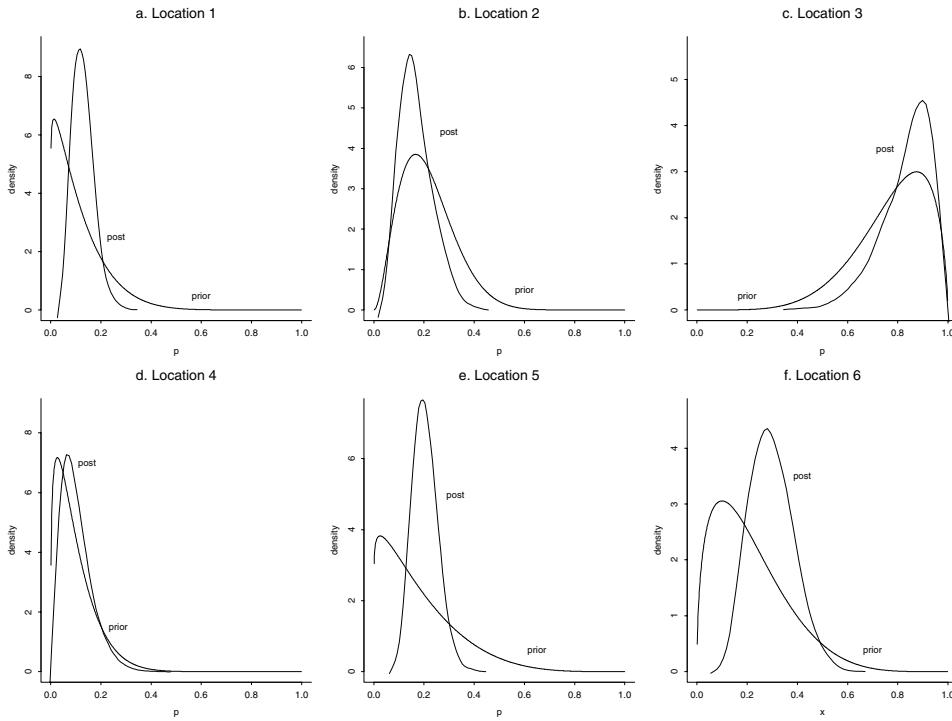
Figure 8.8: Trauma data: Priors and posteriors on  $\tilde{\theta}$ s.

Figure 8.8 gives plots of the priors on the  $\tilde{\theta}_h$ s as well as the posteriors. The priors are generally consistent with the posteriors. Relative to the amount of data, the priors are not overwhelming, being the equivalent of  $57.5 = \sum_{h=1}^6 \tilde{N}_h$  observations compared to 300 data points.

Our initial discussion with Dr. Osler involved eliciting 1st, 50th, and 99th percentiles for each  $\tilde{\theta}_h$ . A Beta distribution involves only two parameters, so these three probability statements over-specify it. There may not exist a Beta distribution that agrees with all three statements. We found Beta distributions that were similar to all three statements and then reconfirmed with Dr. Osler that the Beta distributions did a good job of modeling his prior information.

The first probability  $\tilde{\theta}_1$  corresponds to an individual that “has good physiology, is ‘not bad hurt,’ does not have a lot of reserve,” and for whom there is “added uncertainty due to age.” The Beta(1.1, 8.5) suitably reflects Dr. Osler’s uncertainty about  $\tilde{\theta}_1$ . The median of his prior is around 0.09. The second type of individual “has bad physiology, is very ill, but is young and resilient and is not so bad

hurt." The prior for  $\tilde{\theta}_2$  is Beta(3, 11) with median around 0.20. Incidentally, "bad physiology" and "very ill" apparently refer to bad RTS scores, while how badly hurt one is relates to ISS. The third individual has "bad physiology, a pretty bad injury, and there is much more uncertainty here due to the age factor." The prior is Beta(5.9, 1.7) with median around 0.8. Prior individual four "is young, resilient, and has a big injury." The prior is Beta(1.3, 12) with a median of around 0.07. Dr. Osler had more difficulty with the 5th and 6th types of individuals because their conditions were less extreme than those already considered, and, presumably, because he did not select them. The priors for  $\tilde{\theta}_5$  and  $\tilde{\theta}_6$  are Beta(1.1, 4.9) with approximate median 0.15, and Beta(1.5, 5.5) with approximate median 0.19, respectively.

**EXERCISE 8.8\*** . Use Proposition B.4 in the appendix to derive expression (3). Then show that (3) reduces to (4) in the logistic link case.

**EXERCISE 8.9\***. *Bayes Factors for the O-Ring Data* Below, we give WinBUGS code for calculating the Bayes factor comparing  $M_1$  (logistic link) to  $M_3$  (complementary log-log link) using the O-ring data. You may want to review the method for simulating Bayes factors with parameters of equal dimensions given in Subsection 4.8.3. (a) Before running the code, use (3) to show that the induced prior for  $\beta$  under the complementary log-log transformation is  $p(\beta) \propto \prod_{i=1}^2 \{1 - \exp(-e^{\tilde{x}_i \beta})\}^{a_i-1} \{\exp(-e^{\tilde{x}_i \beta})\}^{b_i} e^{\tilde{x}_i \beta}$ , and thus derive the expression for  $u[i]$  in the code below.

```

model{
  for (i in 1:n){
    y[i] ~ dbin(theta[i],1)
    logit(theta[i]) <- beta[1] + beta[2]*(temp[i]-mt)/sdtemp
  }
  for(i in 1:2){ tildetheta[i] ~ dbeta(a[i],b[i]) }
  beta[1]<-G[1,1]*logit(tildetheta[1])+G[1,2]*logit(tildetheta[2])
  beta[2]<-G[2,1]*logit(tildetheta[1])+G[2,2]*logit(tildetheta[2])
  #Have specified logistic model and induced prior on beta
  #Now specify probs under cloglog model
  for(i in 1:n){
    cloglog(thetastar[i]) <- beta[1]+beta[2]*(temp[i]-mt)/sdtemp
    #Now give terms that go into log lik ratio
    #comparing cloglog in num to logistic in den
    v[i] <- y[i]*(log(thetastar[i])-log(theta[i]))
      +(1-y[i])*(log(1-thetastar[i])-log(1-theta[i]))
  }
  #Finally, give corresponding log of prior ratio
  for(i in 1:2){
    cloglog(tildethetastar[i]) <- beta[1]
      + beta[2]*(ttemp[i]-mt)/sdtemp
    logit(tttildetheta[i]) <- beta[1]
      + beta[2]*(ttemp[i]-mt)/sdtemp
    u[i] <- (a[i]-1)*log(tildethetastar[i])
      - a[i]*log(tttildetheta[i])
      + b[i]*(log(1-tildethetastar[i])
      - log(1-ttildetheta[i]))
      + beta[1]+ beta[2]*(ttemp[i]-mt)/sdtemp
  }
  wstar1 <- sum(v[ ])
  wstar2 <- sum(u[ ])
  w <- exp(wstar1+wstar2)
}

```

```

}
list(tildetheta=c(0.5,0.5))
list(ttemp=c(55,75), mt=69.56522, sdtemp=7.05708,
     n=23, a=c(1.6,1.0), b=c(1.0,1.6))
G[,1]      G[,2]
0.2717391  0.7282609
-0.3528540 0.3528540
END

```

(b) Run the code and obtain  $BF_{31}$ . (c) Give the induced prior on  $\beta$  under the probit model directly using (3). (d) Modify the code and get  $BF_{21}$ . (e) Surmise the value of  $BF_{32}$ .

#### 8.4.2.1 Prior Elicitation

The *who*, *what*, *when*, *where*, *why*, and *how* of prior elicitation:

- *Who*: the expert you are working with.
- *What*: To determine a Beta distribution we need to make two statements about the distribution—two statements because there are two parameters in a Beta distribution. We could specify the mean and variance. We could specify two percentiles, say, the median and the 95th percentile. Most often we choose the mode and a percentile. To this end, some of us developed the Beta-Buster GUI that was discussed in Section 5.1. With the O-ring data, we found two distinct distributions that had the same mode and 66th percentile. But for Beta distributions with both parameters larger than 1, the mode and a percentile uniquely define the Beta distribution.
- *When*: Preferably before data collection, but realistically, any time as long as the prior is elicited *independently* of the data. If you use the current data to obtain the prior, you are using the data twice, which is “cheating.”
- *Where*: At the  $\tilde{x}_h$ s.
- *Why*: The method is more interpretable than direct elicitation for regression coefficients.
- *How*: With considerable hard work. You need to sit down with your expert (if you are your own expert, choose your own posture). Explain exactly what you are asking them to do. Explain percentiles. Explain measures of central tendency. We use the mode because it is easily interpreted as the expert’s best guess. It is a good idea to have the expert over-specify the Beta distribution. While two pieces of information determine a Beta, they may not exhaust the expert’s knowledge. *It is more important to get a prior that accurately describes the expert’s knowledge, than to use a Beta distribution.* But our experience is that we can typically find some Beta distribution that does a good job of approximating an expert’s knowledge. This leads to our ultimate answer to
- *How*: Which is, “repeatedly.” Keep going back to the expert to reconfirm that distributions you have come up with accurately reflect the expert’s knowledge. Many experts are unfamiliar with the process, so their statements may change over short periods of time. Elicit the prior information. Find a distributional model consistent with the prior information. Reconfirm the validity of the distributional model with the expert! When in doubt, allow for more rather than less uncertainty.

By now it should be clear that the process of eliciting a prior is very much a collaboration between the expert and the statistician. The judgement and expertise of both are needed, especially for complex structures like the trauma data. There is no “true” prior, only priors that adequately reflect uncertainty and information.

#### 8.4.2.2 Data Augmentation Priors

Data augmentation priors were first mentioned in Section 2.3. The prior displayed in relation (1) has the same form as the likelihood, so it is a *data augmentation prior (DAP)*. The likelihood times the prior has the form of a likelihood with additional “prior” data  $(\tilde{y}_h, \tilde{N}_h)$ ,  $h = 1, \dots, r$ . In other words,

for the logistic model we can think of the parameters of the prior distribution as a prior sample size  $\tilde{N}_h$  and a prior number of successes  $\tilde{y}_h$  corresponding to the vector of predictors  $\tilde{x}_h$ . With different link functions, different distributions on the  $\tilde{\theta}_h$ s lead to different DAPs. (The likelihood depends on the link function, so DAPs depend on the link function.) We typically use independent Beta priors regardless of the link function, so our priors are only DAPs for logistic models.

When using a DAP, standard frequentist computer programs will find the posterior mode  $\beta_M$  and an asymptotic covariance matrix  $\Sigma(\beta_M)$  for the posterior. From a Bayesian viewpoint, the prior has the same form as the likelihood, so the posterior has the same form as the prior. However, if the posterior has the same form as the likelihood, computer programs designed to maximize the likelihood will also maximize the posterior and perform related computations for the posterior. For the O-ring example, simply augment the observed data with a prior “binomial” observation at 55 degrees consisting of 2.6 trials and 1.6 observed O-ring failures from the Beta(1.6, 1) prior and include a prior observation at 75 degrees with 2.6 trials and 1 O-ring failure from the Beta(1, 1.6) prior. (This assumes that the program will accept non-integer values for the numbers of trials and successes.) The posterior mode of  $\beta$  is the maximum likelihood estimate from the augmented data. The asymptotic covariance matrix computed from the augmented data is the asymptotic dispersion matrix for the normal approximation to the posterior. These quantities are of interest in themselves and can also be useful in creating a good discrete approximation to the posterior. The availability of general programs like WinBUGS make these computations less important.

#### 8.4.2.3 Standardized Variables

Earlier, we mentioned that standardizing the predictor variables can improve the numerical performance of models. Standardizing variables can change both the specification of the prior and some results in the posterior.

For a sample from a single variable  $x$ , with sample mean and standard deviation  $\bar{x}$  and  $s_x$ , we can standardize  $x$  either by looking at  $x - \bar{x}$  or by defining  $z = (x - \bar{x})/s_x$ . With the second standardization, the slope coefficient has a different interpretation due to the different scale. Our focus is on modeling uncertainty about probabilities, so the basic method of inducing a prior on  $\beta$  is unaffected by whether variables are standardized.

We now compare the priors for unstandardized and standardized covariates with the O-ring data. With unstandardized predictors the model was

$$\text{logit}(\theta_i) = \beta_1 + \beta_2 \tilde{x}_i.$$

We let  $\tilde{x}_1 = 55$  and  $\tilde{x}_2 = 75$ . The  $\tilde{\theta}$  vector and the  $\beta$  vector are related through the equation

$$\text{logit} \begin{pmatrix} \tilde{\theta}_1 \\ \tilde{\theta}_2 \end{pmatrix} = \begin{pmatrix} 1 & 55 \\ 1 & 75 \end{pmatrix} \beta \equiv \tilde{X} \beta,$$

with solution

$$\beta = \tilde{X}^{-1} \text{logit}(\tilde{\theta}).$$

With the standardized model using  $z$ , the regression coefficients differ from those in the unstandardized model, so denote them as  $\gamma$ s. The model becomes

$$\text{logit}(\theta_i) = \gamma_1 + \gamma_2 z_i.$$

The mean and standard deviation of temperatures in the data are approximately 69.57 and 7.06, respectively, so the standardized versions of the  $\tilde{x}$ s, call them  $\tilde{z}$ s, are  $-2.06$  and  $0.77$ , respectively. The  $\tilde{\theta}$ s are now defined as

$$\text{logit} \begin{pmatrix} \tilde{\theta}_1 \\ \tilde{\theta}_2 \end{pmatrix} = \begin{pmatrix} 1 & -2.06 \\ 1 & 0.77 \end{pmatrix} \gamma \equiv \tilde{Z} \gamma.$$

We again have two equations in two unknowns and

$$\gamma = \tilde{Z}^{-1} \text{logit}(\tilde{\theta}).$$

**EXERCISE 8.10.** *Prior for the O-Ring Data* The code below is a modification of the code used in Exercise 8.2 for unstandardized variables. Run the code and make inferences about predictive probabilities, the parameters  $\gamma$  and  $\tilde{\theta}$ , the posterior probability that  $\gamma_2 < 0$ , and the odds ratio comparing the odds of O-ring failure at 55 degrees with the odds at 75 degrees. Compare these inferences with those obtained from running suitably modified code with unstandardized covariates from Exercise 8.2. Inferences should be identical (modulo Monte Carlo error), except for  $\beta$  and  $\gamma$ . Find  $\beta$  in terms of  $\gamma$ . Show that  $\Pr(\gamma_2 < 0|y) = \Pr(\beta_2 < 0|y)$  but that the posterior probability of  $\beta_2 < -0.2$  is different from the posterior probability of  $\gamma_2 < -0.2$ .

```
model{
  for(i in 1:n){
    y[i] ~ dbin(theta[i],1)
    logit(theta[i]) <- gamma[1] + gamma[2]*(temp[i]-mt)/sdtemp
  }
  for(i in 1:2){ tildetheta[i] ~ dbeta(a[i],b[i]) }
  gamma[1] <- Ztinv[1,1]*logit(tildetheta[1])
    + Ztinv[1,2]*logit(tildetheta[2])
  gamma[2] <- Ztinv[2,1]*logit(tildetheta[1])
    + Ztinv[2,2]*logit(tildetheta[2])
}
list(tildetheta=c(0.5, 0.5))
list(n=23, mt=69.565, sdtemp=7.057, a=c(1.6,1), b=c(1,1.6))
Ztinv[,1] Ztinv[,2]
0.2717 0.7283
-0.3528 0.3528
END
```

This code includes a  $2 \times 2$  matrix  $Ztinv = \tilde{Z}^{-1}$  used in solving the equations for  $\gamma$  in terms of  $\tilde{\theta}$ . Show that  $Ztinv$  is correct.

**EXERCISE 8.11.** Below we give code for handling the trauma data based on the fully informative prior specified in Table 8.5 with standardized continuous covariates. Below that, we give R code for obtaining  $\tilde{X}^{-1}$  and  $\tilde{Z}^{-1}$ . (a) Modify the code as needed to give a full analysis of the trauma data including assessment of posterior probabilities that regression coefficients are positive (or negative), estimates of probabilities of “death on the table” for the 16 possible combinations of (*ISS, RTS, AGE, TI*) corresponding to *ISS* = 20, 40, *RTS* = 3.34, 5.74, *AGE* = 10, 60, and *TI* = 0, 1. As part of your analysis, create a table of entries that includes the median, and a 95% probability interval for each combination. Inferences are for the proportions of deaths in the populations of trauma patients that fall into these 16 categories. Compare with results obtained in Figure 8.5.

```
model{
  for(i in 1:n){
    death[i] ~ dbern(theta[i])
    logit(theta[i]) <- gamma[1] + gamma[2]*(ISS[i]-14.3)/11
      + gamma[3]*(RTS[i]-7.29)/1.25
      + gamma[4]*(AGE[i]-31.4)/17
      + gamma[5]*TI[i]
      + gamma[6]*(AGE[i]-31.4)*TI[i]/17
  }
  for(i in 1:6){
```

```

tildetheta[i] ~ dbeta(a[i],b[i])
v[i] <- log(tildetheta[i]/(1-tildetheta[i]))
gamma[i] <- inprod(Ztinv[i,1:6], v[1:6])
}
junk <- ID[1]
}
list(tildetheta=c(0.5,0.5,0.5,0.5,0.5,0.5))
list(n=300, a=c(1.1,3,5.9,1.3,1.1,1.5), b=c(8.5,11,1.7,12,4.9,5.5))
Ztinv[,1] Ztinv[,2] Ztinv[,3] Ztinv[,4] Ztinv[,5] Ztinv[,6]
-2.603   -2.460   -3.702   -3.702   6.063   7.403
-0.343   -0.343   0.343   0.343   0.686   -0.686
-2.091   -2.091   -2.091   -2.091   4.182   4.182
2.901    2.218   2.559   2.559   -5.119   -5.119
1.149    1.005   1.005   1.149   -3.154   -1.154
-5.460   -4.778   -4.778   -5.460   10.238  10.238
END
ID[ ]   death[ ]   ISS[ ]   TI[ ]   RTS[ ]   AGE[ ]
2979     0          1          1      7.8408   25
1167     0          9          0      7.8408   63
116      0         29          0      2.9304   32
remaining 297 data lines go here
END

```

(b) Place  $U[0,1]$  priors on  $\tilde{\theta}_4$ ,  $\tilde{\theta}_5$ , and  $\tilde{\theta}_6$  leaving the other prior information the same and rerun the code. Comment on any major changes in the analysis from part (a). (c) Using the full prior from Table 8.5, modify the code to handle untransformed covariates. Compare your estimates of the 16 probabilities from part (a); they should be nearly the same if everything was done correctly since this is just a reparameterization of the model used there. Is there any noticeable change in the numerical performance of WinBUGS? You will have to invert the  $\tilde{X}$  matrix, called `Xtilde` in the R code below, to replace `Ztinv` in the WinBUGS code.

We now discuss how to get the matrix  $\tilde{Z}^{-1}$  where  $\tilde{Z}$  is  $\tilde{X}$  modified so that the continuous covariates ISS, RTS, and AGE are standardized. The process starts by converting the data in Table 8.5 to the object `Xtilde` below. This is copied into a new script file in R as a  $36 \times 1$  row vector. Use the `matrix` command to convert `Xtilde` to a  $6 \times 6$  matrix, `Xtilde[1:6,1:6]`. Type the name `Xtilde` to see these results. Using the means and standard deviations of ISS, RTS, and AGE, we defined a new standardized matrix, `Xtilde[1:6,1:6]`, by taking the first new column to be the old first column, the next three columns are the standardized old columns, the fifth column is the same as before, and the sixth column is the element-by-element product of the current fourth and fifth columns. Then invert `Xtilde` to get `Ztinv`, which we copy (and edit), and place at the end of our WinBUGS program for use in analyzing the trauma data based on standardized continuous covariates.

```

# R code for generating Ztinv
Xtilde <- c(1,25,7.84,60,0,0,
           1,25,3.34,10,0,0,
           1,41,3.34,60,1,60,
           1,41,7.84,10,1,10,
           1,33,5.74,35,0,0,
           1,33,5.74,35,1,35)
Xtilde <- matrix(Xtilde, nrow=6, ncol=6, byrow=TRUE)
Ztinv <- solve(Xtilde)
m.iss <- mean(ISS)

```

```

sd.iss <- sd(ISS)
m.rts <- mean(RTS)
sd.rts <- sd(RTS)
m.age <- mean(AGE)
sd.age <- sd(AGE)
Xtilde <- matrix(0,6,6)
Xtilde[,1] = Xtilde[,1]
Xtilde[,2] = (Xtilde[,2]-m.iss)/sd.iss # ISS mean 14.28 sd 10.98
Xtilde[,3] = (Xtilde[,3]-m.rts)/sd.rts # RTS mean 7.287 sd 1.25
Xtilde[,4] = (Xtilde[,4]-m.age)/sd.age # AGE mean 31.4 sd 17.06
Xtilde[,5] = Xtilde[,5] # TI 0/1 variate
Xtilde[,6] = Xtilde[,4]*Xtilde[,5] # Interaction
Ztinv <- solve(Xtilde)
round(Ztinv,3)

```

#### 8.4.3 Reference Priors

Our general model has  $r$  predictor variables. We elicited expert knowledge about  $r$  probabilities  $\tilde{\theta}_h$  and induced a prior on  $\beta$ . When  $r$  is large, this becomes difficult. Our trauma illustration with  $r = 6$  involved considerable effort in interchanges between statisticians and an expert. This approach may not be practical for larger values of  $r$ . Moreover, if the sample size  $n$  is large, it may not be important to develop informative priors since the data should dominate the prior. In this subsection, we discuss reference priors that involve no actual prior information about the parameters even though they make statements about the parameters. In the next subsection we discuss partially informative priors that involve combining expert knowledge about fewer than  $r$  values  $\tilde{\theta}_h$  along with a reference prior. In our experience, it rarely does harm to elicit some real prior information and even a partially informative prior is often quite helpful.

In Subsection 5.1.3 we discussed reference priors for binomial samples. For logistic regression, just as there are two approaches to defining an informative prior, i.e., directly on  $\beta$  or inducing it from  $\tilde{\theta}_h$ s, there are two approaches to defining a reference prior.

The BCJ method requires specifying  $\tilde{x}_h$ ,  $h = 1, \dots, r$ . If we are willing to specify the  $\tilde{x}_h$ s but not willing to specify informative priors for the corresponding  $\tilde{\theta}_h$ s, we can put independent proper reference priors on the  $\tilde{\theta}_h$ s and induce a proper reference prior on the  $\beta_j$ s. As discussed in Subsection 5.1.1, the proper reference priors for probabilities are Jeffreys' and the uniform.

Historically, the improper flat prior  $p(\beta) = 1$  has been used as a reference prior for regression coefficients. As in Example 4.6.3, this corresponds for any  $\tilde{\theta}_h$  to essentially splitting the prior probability evenly between 0 and 1 (see Exercise 8.12). The flat prior usually has little effect on the posterior distribution.

The traditional method for defining an informative prior directly on  $\beta$  uses a multivariate normal distribution. Multivariate normals are now also used as proper reference priors. To this end a common practice is to use independent  $\beta_j \sim N(0, b)$  distributions where  $b$  is large, often  $b = 10^6$ . This prior induces priors on any  $\tilde{\theta}_h$ s that put half the prior probability very near 0 and the other half very near 1. The reference priors put most of their probability on regression coefficients that are very far from 0, which ends up putting most of the prior probability on values of the log-odds that are near  $\pm\infty$ , and that correspond to probabilities of 0 and 1. As silly as these priors seem, they will usually not affect the posterior very much.

We suggest using independent normal priors but picking  $b$  so that the induced distributions on the  $\theta_i$ s are as close to uniform as we can make them. Specifically, we suggest independent  $\beta_j \sim N(0, s_j^2/b)$  priors, where  $s_j$  is 1 if  $x_j$  is dichotomous and, if  $x_j$  is continuous,  $s_j$  is the sample standard deviation. Of course if  $x_j$  has already been standardized,  $s_j^2 = 1$ . We want all of the continuous regression coefficients on the same scale, and all of the dichotomous ones on their own scale, before looking for an appropriate  $b$ . With  $\theta_i$ s determined by  $\text{logit}(\theta_i) = x'_i \beta$ ,  $i = 1, \dots, n$ , we select  $b$  to make

the prior distributions on the  $\theta_i$ s as uniform as possible. *The other methods of defining reference priors are pretty straightforward, so most of our discussion in this subsection and the next is devoted to this suggestion.*

**EXERCISE 8.12 One-sample Problem.** In case you have forgotten Chapter 4, let  $y|\theta \sim \text{Bin}(n, \theta)$ , and let  $\theta = e^\beta / (1 + e^\beta)$ . Let  $p(\beta) = 1$  be an improper flat prior on  $\beta$ . Then use Proposition B.4 to obtain the induced improper prior  $p(\theta) \propto \theta^{-1}(1 - \theta)^{-1}$ , which has infinite density at 0 and at 1.

**EXERCISE 8.13. Two-sample Problem.** Consider two independent Binomial samples with probabilities  $\theta_1$  and  $\theta_2$  as discussed in Example 8.4.1. Define  $\beta_1 = \text{logit}(\theta_1)$  and  $\beta_2 = \text{logit}(\theta_2) - \text{logit}(\theta_1)$ . (a) For independent  $\beta_j \sim N(0, 10^3)$ ,  $j = 1, 2$ , find the prior distributions on  $\theta_1$  and  $\theta_2$ . The following WinBUGS code gives the induced prior on  $\theta = \tilde{\theta}$ .

```
model{
  beta[1] ~ dnorm(0,0.001)
  beta[2] ~ dnorm(0,0.001)
  tildetheta[1]<- exp(beta[1])/(1+exp(beta[1]))
  tildetheta[2]<- exp(beta[1] + beta[2])/(1+exp(beta[1] + beta[2]))
}
```

Note that the  $\tilde{\theta}_j$ s prior densities are concentrated near 0 and 1. (b) What happens when  $\text{dnorm}(0, 0.001)$  is replaced by  $\text{dflat}()$ ? (c) Place independent  $N(0, 1/b)$  priors on the  $\beta_j$ s and find a value of  $b$  that induces reasonably spread out (ideally, uniform) priors on the  $\tilde{\theta}_j$ s. (d) Now let  $\tilde{\theta}_j \sim U[0, 1]$ . The following code will induce a prior on the  $\beta_j$ s.

```
model{
  for(j in 1:2){ tildetheta[j] ~ dunif(0,1) }
  beta[1] <- logit(tildetheta[1])
  beta[2] <- logit(tildetheta[2]) - logit(tildetheta[1])
}
```

Run the code and look at pictures of the induced priors on the  $\beta_j$ s.

**EXERCISE 8.14.** The WinBUGS code below specifies a reference prior on  $\beta$  for the O-ring data. The predictor variable is standardized and the priors on the components of  $\beta$  are independent  $N(0, 1/b)$ s. The code needs to be modified to read the data, cf. Exercise 8.2. The code includes a model for the O-ring failures  $y$  that has nothing to do with temperature. This was done to comply with the WinBUGS requirement of using all the variables in a data set; we would get an error message if  $y$  were not used in the program in some way. Alternatively, we could have modified the data set to only include temperatures (no response data  $y$ ).

(a) Find a reasonable choice of  $b$  for the O-ring data. Do this by plotting the induced probabilities of O-ring failure for all 23 temperatures in the data, for various choices of  $b$ . Pick a value of  $b$  that gives the most uniform induced priors on the  $\theta$ s. (b) After selecting a value of  $b$ , modify the code below to use this prior in an analysis of the O-ring data. Compare the posterior analysis with that based on the choice  $b = 0.000001$ , and with a third choice,  $b = 0.25$ . Comment on convergence issues, and compare estimates of the probability of O-ring failure at 55 and 75 degrees for each analysis.

```
model{
  for(j in 1:2){ beta[j] ~ dnorm(0,b) }
  for(i in 1:23){
    y[i] ~ dbin(0.5,1)
    logit(theta[i]) <- beta[1] + beta[2]*(temp[i]-mt)/sdtemp
  }
}
```

```
list(beta = c(0,0))
list(mt=69.565, sdtemp=7.057, b=1)
```

This approach would be odd for the two-sample binomial problem since we could just model  $\theta_1$  and  $\theta_2$  with independent uniform priors directly, rather than taking the circular route of trying to find a precision  $b$  so that independent normal priors on  $\beta_1$  and  $\beta_2$  induce distributions on  $\theta_1$  and  $\theta_2$  that are nearly uniform. Nonetheless, we explore its role as the simplest case of binomial regression.

EXERCISE 8.15. Find a reasonable choice of  $b$  for the two-sample problem.

#### 8.4.4 Partial Prior Information

Specifying a prior distribution for a regression with several predictor variables can be a lot of work. We now consider specifying a prior distribution with partial information in which some aspects of the prior correspond to expert knowledge while other aspects use reference priors. Specifying partial information involves specifying a prior for fewer parameters  $\tilde{\theta}_h$  than the number of regression parameters, hence reducing our prior elicitation workload. There are two approaches depending on whether we are willing to specify all of the  $\tilde{x}_h$ s,  $h = 1, \dots, r$ . If we are willing to specify them all, then put informative priors on some  $\tilde{\theta}_h$ s and reference priors on the rest. Most of the discussion examines the case where we specify fewer than  $r$  vectors  $\tilde{x}_h$ . BCJ (1996, 1997) give an alternative approach to partial prior information.

With fewer than  $r$  vectors, we choose the  $\tilde{x}_h$ s so that the corresponding  $\tilde{\theta}_h$ s only induce information on a subset of the regression coefficients. To do this, the  $\tilde{x}_h$ s are chosen so that the predictor variables associated with the remaining coefficients do not vary. (The next subsection discusses some theoretical considerations behind our approach.) We then place independent reference distributions on these remaining regression coefficients. The reference distributions can be the improper flat priors  $p(\beta_j) = 1$  or they can be proper reference distributions  $N(0, 1/b)$  with  $b$  a small number like 0.001 or  $10^{-6}$  or, with standardized predictors,  $b$  chosen to make the priors on the  $\theta_i$ s close to uniform.

EXAMPLE 8.4.1 CONTINUED. *Two-samples.* Again assume that  $\tilde{\theta}_1 \equiv \theta_1 \sim \text{Beta}(11.26, 11.26)$  but that there is now no information for  $\tilde{\theta}_2 \equiv \theta_2$ . In the regression parameterization of two samples,  $\beta_1 = \text{logit}(\tilde{\theta}_1)$ , so the informative prior on  $\theta_1$  induces an informative prior on  $\beta_1$ . The other regression parameter is  $\beta_2 = \text{logit}(\tilde{\theta}_2) - \text{logit}(\tilde{\theta}_1)$ . At this point, we have two choices. We can put a reference prior on  $\theta_2$ , such as a uniform or Jeffreys' prior, and induce a prior on  $\beta_2$ . Alternatively, we can put a reference prior directly on  $\beta_2$ .

It seems reasonable to assume that information about  $\beta_2$ , which is the log-odds ratio of international recruitment to domestic recruitment, is independent of  $\beta_1$ , especially since we claim to be ignorant of  $\beta_2$ . We could choose any reference prior for  $\beta_2$  but we like  $\beta_2 \sim N(0, 1/b)$ .

EXERCISE 8.16. For the two-sample problem, the following WinBUGS code induces a prior on  $\beta$  based on the partially informative specification.

```
model{
  tildetheta[1] ~ dbeta(11.26,11.26)
  beta[1] <- logit(tildetheta[1])
  beta[2] ~ dnorm(0,b)
  tildetheta[2] <- exp(beta[1]+beta[2])/(1+exp(beta[1]+beta[2]))
}
list(b=1)
```

Plot the induced distribution on  $\tilde{\theta}_2$  for various values of  $b$ . Choose a value  $b$  and analyze the data. Revise the code so that  $\tilde{\theta}_2$  has a  $U[0, 1]$  prior, reanalyze the data, and compare results. Is there an advantage to using the uniform distribution on  $\tilde{\theta}_2$ ?

EXAMPLE 8.4.4. *O-Ring Data.* Previously, we specified  $\tilde{x}_1 = 55$  and  $\tilde{x}_2 = 75$  and gave informative priors for  $\tilde{\theta}_1$  and  $\tilde{\theta}_2$ . We could have specified a prior on only one  $\tilde{\theta}_h$  and put a reference prior on the other.

Now consider the case where we only specify one  $\tilde{x}_h$ . The standard model for the data is  $\text{logit}(\theta_i) = \beta_1 + \beta_2 \tilde{x}_i$ ,  $i = 1, \dots, 23$ . The intercept  $\beta_1$  is the log-odds of O-ring failure when the temperature is zero degrees Fahrenheit. It is difficult to have good prior knowledge about a parameter that is so far out of the range of normal experience. Instead, we standardize the model to

$$\text{logit}(\theta_i) = \gamma_1 + \gamma_2 (\tilde{x}_i - \bar{x}_.) / s_x.$$

Here,  $\gamma_1$  is quite an interesting parameter; it is the log-odds of O-ring failure at  $\bar{x}_.$ , the mean temperature of the 23 actual flights.

Suppose we only have expert input for a single  $\tilde{\theta}$ , namely, the probability of O-ring failure at  $\bar{x}_.$ . Essentially, we are asking for expert knowledge about the probability of O-ring failure at common operating conditions, e.g., baseline conditions. In any case, we have expert knowledge on  $\gamma_1 = \text{logit}(\tilde{\theta})$ . This knowledge may be independent of the expert's knowledge about how temperature affects O-ring performance. We could use any reference prior on  $\gamma_2$  but we typically choose an independent  $N(0, 1/b)$  prior.

Suppose the expert is most comfortable with, say, the probability of O-ring failure at the relatively low temperature of 55 degrees, and only provides information about that. If the expert has no idea whether the probability will go up or down or stay the same as temperature is increased, our approach works just fine. On the other hand, the expert may know that the probability decreases as temperature increases, which is equivalent to believing that  $\gamma_2 < 0$ . If the expert has no real idea of the magnitude of  $\gamma_2$  beyond its being negative, it would be reasonable to assume independence of  $\gamma_1$  and  $\gamma_2$  subject to this constraint. The prior on  $\gamma_2$  could be a *half-normal distribution*, i.e., a  $N(0, 1/b)$  conditional on  $\gamma_2 < 0$ . The reader can try different values of  $b$  to obtain prior distributions for the  $\theta_i$ s that are sufficiently spread out without too much probability near 0.

If the expert was not 100% certain that  $\gamma_2 < 0$ , a  $N(a, 1/b)$  prior can be placed on  $\gamma_2$  that has, say, 90th percentile equal to 0. A second percentile could be chosen with, say, 5th percentile equal to  $-2$ . Such a prior would satisfy  $1.28 = -a\sqrt{b}$  and  $-1.645 = (-2 - a)\sqrt{b}$ . Solving gives  $a = -2.56/2.925, \sqrt{b} = 2.925/2$ .

EXERCISE 8.17. *O-Ring Data.* Construct several partially informative priors for the O-ring data, analyze the data, and compare the results. (a) Using the prior information given in Example 8.4.2, construct a reasonable prior for the probability of O-ring failure at the average temperature of 69.6 degrees. Justify your selection. Then select a  $N(0, 1/b)$  prior for  $\beta_2$  that results in induced priors on probabilities of failure at the lowest and the highest temperatures, i.e., temperatures that are not too close to 70 degrees, that are as uniform as possible. (b) With your choice of  $b$ , obtain inferences. Then let  $b = 0.001$  and reanalyze the data. Compare results. (c) Run the code with  $\tilde{x}_1 = 55, \tilde{x}_2 = 75, \tilde{\theta}_1 \sim \text{Beta}(1.6, 1)$ , and  $\tilde{\theta}_2 \sim \text{Beta}(1, 1)$ . Also run the code with  $\tilde{\theta}_1 \sim \text{Beta}(0.5, 0.5)$  and  $\tilde{\theta}_2 \sim \text{Beta}(1, 1.6)$ . Compare all results.

EXAMPLE 8.4.3 CONTINUED. *Trauma Data.* The data model is, for  $i = 1, \dots, 300$ ,  $y_i | \theta_i \sim \text{Bern}(\theta_i)$ ,

$$\text{logit}(\theta_i) = \beta_1 + \beta_2 \text{ISS}_i + \beta_3 \text{RTS}_i + \beta_4 \text{AGE}_i + \beta_5 \text{TI}_i + \beta_6 (\text{AGE} * \text{TI})_i,$$

which can be rewritten

$$\text{logit}(\theta_i) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i4} x_{i5}$$

with standardized version (only standardizing continuous covariates)

$$\begin{aligned} \text{logit}(\theta_i) &= \gamma_1 + \gamma_2 (x_{i2} - \bar{x}_{.2}) / s_2 + \gamma_3 (x_{i3} - \bar{x}_{.3}) / s_3 \\ &\quad + \gamma_4 (x_{i4} - \bar{x}_{.4}) / s_4 + \gamma_5 x_{i5} + \gamma_6 (x_{i4} - \bar{x}_{.4}) x_{i5} / s_4. \end{aligned}$$

We work with the standardized version.

Table 8.6: Trauma data: Partial prior specification (new parametrization).

$i$	Design for Prior						Beta( $\tilde{y}$ , $\bar{N} - \tilde{y}$ )		
	$\tilde{x}'_i$						$\tilde{y}$	$\bar{N} - \tilde{y}$	Prior Mode
1	1	0	0	0	0	0	2.06	21.2	0.05
2	1	1	0	0	0	0	2.70	10.66	0.15
3	1	1	-2	0	0	0	1.9	3.7	0.25

Let's assume that our expert was only willing to specify three  $\tilde{x}_i$ 's and three probabilities. We take as "baseline" conditions a patient who has the average values of ISS ( $\bar{x}_{.2} = 14.3$ ), RTS ( $\bar{x}_{.3} = 7.29$ ), and AGE ( $\bar{x}_{.4} = 31.4$ ), and has a blunt injury (TI = 0). We indirectly specify prior information on the three parameters  $\gamma_1$ ,  $\gamma_2$ , and  $\gamma_3$ . Let  $\tilde{x}_1$  correspond to the baseline conditions, so  $\tilde{\theta}_1$  corresponds to  $\text{logit}(\tilde{\theta}_1) = \gamma_1$ . We let  $\tilde{x}_2$  correspond to a baseline individual except with an ISS score that is one standard deviation *above* the mean, i.e., the second component of  $\tilde{x}_2$  is  $\tilde{x}_{22} = \bar{x}_{.2} + s_2 = 14.3 + 11 = 25.3$ , so that  $\text{logit}(\tilde{\theta}_2) = \gamma_1 + \gamma_2$ . Finally, let  $\tilde{x}_3$  be like  $\tilde{x}_2$  except with an RTS that is two standard deviations *below* the mean, i.e.,  $\tilde{x}_{33} = \bar{x}_{.3} - 2s_3 = 7.29 - 2(1.25) = 4.79$ . It follows that  $\text{logit}(\tilde{\theta}_3) = \gamma_1 + \gamma_2 - 2\gamma_3$ .

Clearly, the prior information involves only  $\gamma_1$ ,  $\gamma_2$ , and  $\gamma_3$ . More importantly, since AGE, TI, and (AGE \* TI) never varied from their baseline values, it is reasonable to treat the information on  $\gamma_1$ ,  $\gamma_2$ , and  $\gamma_3$  as independent of  $\gamma_4$ ,  $\gamma_5$ , and  $\gamma_6$ , since these parameters only affect how probabilities *change* as AGE, TI, and (AGE \* TI) change from the baseline. Finally, we complete the prior by taking independent  $\gamma_j \sim N(0, 1/b)$  distributions for  $j = 4, 5, 6$ . As in the simple regression examples,  $b$  is determined either as an arbitrary small number or to give more uncertainty for  $\theta_i$ 's in the data than exists in the elicited prior distributions for the  $\tilde{\theta}_i$ 's but also to avoid mass piling up at 0 and 1, e.g., uniform. Table 8.6 summarizes the partial prior specification.

*Remark:* It might be tempting to think that since the  $\tilde{\theta}_i$ 's do not depend on  $\gamma_4$ ,  $\gamma_5$ , and  $\gamma_6$ , that the independence assumption is no longer important. The  $\theta_i$ 's may determine a distribution on  $\gamma_1$ ,  $\gamma_2$ , and  $\gamma_3$  alone, but we still need to worry about the conditional distribution of  $\gamma_4$ ,  $\gamma_5$ , and  $\gamma_6$  given  $\gamma_1$ ,  $\gamma_2$ , and  $\gamma_3$ . It is the use of a baseline model that justifies independence. We discuss this further in the next subsection.

**EXERCISE 8.18.** In this exercise we place a partially informative prior on the regression coefficients for the trauma data. We don't have further access to an expert, so we used the prior information from Dr. Osler that was given in Table 8.5 to construct a prior on three  $\tilde{\theta}_i$ 's. These priors are given in Table 8.6 using the standardized variables. Below is WinBUGS code for selecting a partially informative prior. It does not use the data, other than the covariate information. Run the code with different values of  $b$  and monitor the  $\theta_i$ 's. Pick the value of  $b$  that seems to be the *least informative*. (This is somewhat subjective.) Some of the priors on  $\theta_i$ 's must be informative, since some  $\theta_i$ 's are located near a  $\tilde{\theta}_i$ , so focus on  $\theta_i$ 's that have predictor combinations dissimilar to those given in Table 8.6.

```
model{
  for(i in 1:n){
    death[i] ~ dbern(0.5)
    logit(theta[i]) <- gamma[1] + gamma[2]*(ISS[i]-14.3)/11
      + gamma[3]*(RTS[i] - 7.29)/1.25
      + gamma[4]*(AGE[i] - 31.4)/17
      + gamma[5]*TI[i]
      + gamma[6]*(AGE[i]-31.4)*TI[i]/17
  }
}
```

```

for(i in 1:3){ tildetheta[i] ~ dbeta(a[i],bb[i]) }
gamma[1] <- logit(tildetheta[1])
gamma[2] <- logit(tildetheta[2]) - gamma[1]
gamma[3] <- (logit(tildetheta[2]) - logit(tildetheta[3]))/2
for(j in 4:6){ gamma[j] ~ dnorm(0,b) }
junk <- ID[1]
}
list(tildetheta = c(0.1,0.1,0.1))
list(n=300, a=c(2.06,2.7,1.9), bb=c(21.2,10.66,3.7), b=1)
ID[ ] death[ ] ISS[ ] TI[ ] RTS[ ] AGE[ ]
2979      0        1      1    7.8408     25
1167      0        9      0    7.8408     63
116      0       29      0    2.9304     32
remaining 297 data lines go here
END

```

**EXERCISE 8.19.** Analyze the trauma data by modifying the code from the previous exercise so that you can make inferences about the 16 particular probabilities of death discussed in Exercise 8.11. (a) Run the code with your  $b$  from (i) the previous exercise, (ii) using the prior `dflat()` for all of the  $\beta_j$ s, and (iii) using `dflat()` for  $\beta_4, \beta_5$ , and  $\beta_6$ . Compare results for the regression coefficients and for the 16 probabilities. (b) Modify the code from Exercise 8.11, where a fully informative prior was used with standardized ISS, RTS, and AGE, to incorporate three  $U[0, 1]$  priors corresponding to  $\tilde{x}_h$ ,  $h = 4, 5, 6$ , listed in Table 8.5. The first 3  $\tilde{\theta}_h$ s will have the same informative Beta priors specified there. Evaluate the impact of the modified prior on the analysis.

#### 8.4.5 Partial Priors: Theoretical Considerations\*

As discussed in Subsection 8.4.2, we specify full prior information by specifying an  $r \times r$  covariate matrix  $\tilde{X}$  and a prior distribution for the corresponding vector of success probabilities,  $\tilde{\theta} = F(\tilde{X}\beta)$ . As mentioned in the previous subsection, if  $\tilde{X}$  is fully specified, partial information can take the form of real prior information on some components of  $\tilde{\theta}$  and proper reference priors on other components. In any case, the distribution on  $\tilde{\theta}$  determines the distribution on  $\beta$  via

$$\beta = \tilde{X}^{-1}F^{-1}(\tilde{\theta}).$$

This form is particularly nice for computer simulations because we sample from the distribution of  $\tilde{\theta}$ , and apply the function to obtain a sample from  $\beta$ .

For our other form of partial prior information, we partition  $\beta$  into subvectors with  $r_1$  and  $r_2 = r - r_1$  components and define  $\beta' = (\beta'_1, \beta'_2)$ . We use a reference prior on  $\beta_2$  and use our partial prior information to determine the distribution of  $\beta_1$  given  $\beta_2$ , which completes the specification of a joint distribution on  $\beta_1$  and  $\beta_2$ . Rather than specifying the distribution of  $\beta_1|\beta_2$  directly, for a vector of probabilities  $\tilde{\theta}$  and functions  $g_1$  and  $g_2$  we write  $\beta_1 = g_1(\tilde{\theta}) + g_2(\beta_2)$ . We assume that  $\tilde{\theta} \perp\!\!\!\perp g_2(\beta_2)$  and specify the marginal distribution of  $\tilde{\theta}$ . Thus the marginal distributions of  $\tilde{\theta}$  and  $\beta_2$  completely determine the joint distribution on  $\beta_1$  and  $\beta_2$ .

Our partial prior information takes the form of specifying an  $r_1 \times r$  matrix  $\tilde{X}$ , which we partition into  $r_1 \times r_1$  and  $r_1 \times r_2$  submatrices,

$$\tilde{X} = [\tilde{X}_1 \quad \tilde{X}_2].$$

$\tilde{\theta}$  is the  $r_1$  vector of success probabilities that corresponds to  $\tilde{X}$ . The model gives  $\tilde{\theta} = F(\tilde{X}\beta) = F(\tilde{X}_1\beta_1 + \tilde{X}_2\beta_2)$ . Solving for  $\beta_1$ ,

$$\beta_1 = \tilde{X}_1^{-1} [F^{-1}(\tilde{\theta}) - \tilde{X}_2\beta_2] = \tilde{X}_1^{-1}F^{-1}(\tilde{\theta}) - \tilde{X}_1^{-1}\tilde{X}_2\beta_2.$$

In constructing  $\tilde{X}_2$ , we fix each row vector to be identical so that we are holding the corresponding  $r_2$  predictor variables associated with  $\beta_2$  fixed for all  $r_1$  specifications that we intend to make. We do this so that we think only about how the variables associated with  $\beta_1$  affect the corresponding  $r_1$  probabilities while the remaining variables are fixed. The particular values that we fix are denoted  $\tilde{x}_{2*}$ . So we have  $\tilde{X}_2 = J\tilde{x}'_{2*}$  where  $J$  is an  $r_1 \times 1$  vector of ones. Moreover, we assume that the model contains an intercept and that the first column of  $\tilde{X}_1$  is a column of 1s. It follows that  $\tilde{X}_1^{-1}J = (1, 0, \dots, 0)'$ , and thus  $\tilde{X}_1^{-1}\tilde{X}_2\beta_2 = \tilde{X}_1^{-1}J\tilde{x}'_{2*}\beta_2 = (1, 0, \dots, 0)'\tilde{x}'_{2*}\beta_2$ , which leads to

$$\beta_1 = \tilde{X}_1^{-1}F^{-1}(\tilde{\theta}) - \begin{bmatrix} \tilde{x}'_{2*}\beta_2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

We then place an informative prior on  $\tilde{\theta}$  that does not depend on  $\tilde{x}'_{2*}\beta_2$ , so we are assuming that  $\tilde{\theta}$  is independent of  $\tilde{x}'_{2*}\beta_2$ . The reasonableness of this assumption depends on our choice for  $\tilde{x}_{2*}$ . The reference prior on  $\beta_2$  along with the informative prior on  $\tilde{\theta}$  determines the joint prior on  $\beta_1$  and  $\beta_2$ .

With independent distributions for  $\tilde{\theta}$  and  $\beta_2$ , independent Beta distributions for  $\tilde{\theta}$ s, and a reference normal distribution for  $\beta_2$ , it is a simple matter to sample from  $\beta = (\beta'_1, \beta'_2)'$ . Moreover, this relationship makes for easy sampling of the posterior.

A natural choice for  $\tilde{x}_{2*}$  is the mean of the corresponding variables in the data sample. Mathematically, the simplest choice for  $\tilde{x}_{2*}$  is the 0 vector. If we standardize continuous predictor variables and set any categorical variables to zero “reference” values, the natural choice and the simplest choice are the same. Moreover, the key assumption that  $\tilde{\theta} \perp\!\!\!\perp \tilde{x}'_{2*}\beta_2$  is probably most reasonable when  $\tilde{x}_{2*}$  is the mean vector. When  $\tilde{x}'_{2*}\beta_2 = 0$ , which it generally is in our examples, then the independence assumption is automatically satisfied since everything is independent of the constant 0.

For the components of  $\beta_2$ , we can use any reference distribution including independent  $\beta_j \sim N(0, 1/b)$  distributions where  $j = r_1 + 1, \dots, r$ . If we assume that the continuous covariates have been standardized, the precision factor  $b$  can be chosen large enough to avoid mass piling up at 0 and 1 for predictor values in the data but small enough to be less precise than the actually elicited prior information.

## 8.5 Mixed Models

We now introduce the ideas of Section 4.12 on hierarchical models into logistic regression. Specifically, this takes the form of allowing some regression coefficients to be random in the sampling model. Two primary reasons for introducing random effects models are that they can introduce correlation among Bernoulli observations when appropriate and that they can adjust for missing or unknown covariates/predictors that are common to a group of observations while not fundamentally changing the nature of the statistical inference. Readers with less statistical background might want to postpone this section until after developing more familiarity with regression in Chapter 9.

Suppose we sample 100 hospitals and then randomly sample 20 patients from each hospital. At the end of each patient’s (non-life-threatening) stay, we ask them if they were satisfied with their care in the hospital (our binary response). In addition, we track certain hospital characteristics like the number of nursing staff per patient, the hospital’s average number of patient contacts with doctors, etc. We might even track patient-specific items like their age and their specific disease or hospital unit admission. But, no matter how much covariate information is collected, there will always be variables that were omitted because they were too expensive to measure or because they were not considered important. To the extent that these are hospital variables and not patient variables, we can accommodate them by incorporating an overall hospital effect.

Let  $x_{ij}$  denote the vector of covariate information for the  $j$ th patient in the  $i$ th hospital (excluding an intercept), let  $y_{ij}$  be the indicator of the satisfaction of patient  $j$  in hospital  $i$ , and let  $\theta_{ij} = \Pr(y_{ij} = 1 | x_{ij})$ . The model

$$\text{logit}(\theta_{ij}) = \mu + x'_{ij}\beta \tag{1}$$

allows for the effects of measured hospital and patient characteristics on patient satisfaction. It allows us to estimate a patient's satisfaction for any hospital in the data and to predict patient satisfaction for any new hospital on which we have the appropriate covariates.

If we suspect that important hospital variables have been omitted, a natural surrogate for them is to add an effect for each hospital in model (1), namely

$$\text{logit}(\theta_{ij}) = x'_{ij}\beta + \gamma_i.$$

If  $\gamma_i$  is treated as a fixed effect, the model does not allow us to discuss hospitals in general. Every hospital has its own unique effect. There is no possibility of predicting patient satisfaction for a new hospital because there is no way to estimate  $\gamma_i$  for a new hospital. Moreover, the hospital variables incorporated in  $x_{ij}$  would be redundant, serving no purpose in the model. (Hospital variables are elements of the vector  $x_{ij}$  that *never* vary with  $j$ .) However, if we assume that the  $\gamma_i$ s are random effects, e.g.,

$$\gamma_i | \mu, \tau \stackrel{iid}{\sim} N(\mu, 1/\tau),$$

each hospital is sampled from a population of hospitals and we can discuss both the mean and variability of this population. The satisfaction probabilities for patients in each hospital go up or down depending on the sign of  $\gamma_i - \mu$ . A hospital with  $\gamma_i = \mu$  is a "typical" hospital, which allows us to make point predictions of the probability of satisfaction for a patient with covariates  $x$  from a new hospital. Moreover, the variability in the  $\gamma_i$ s affects the variability in the probabilities of satisfaction for patients from a new hospital. Finally, this model also correlates responses within each hospital, i.e.,  $\{y_{ij} : j = 1, \dots, 20\}$  are all correlated, while responses from different hospitals are independent. The fact that all patients from hospital  $i$  share the same random  $\gamma_i$  effect induces the correlation.

Our usual binomial regression sampling model is for  $k = 1, \dots, n$ ,

$$\begin{aligned} y_k | \theta_k &\stackrel{ind}{\sim} \text{Bin}(N_k, \theta_k) \\ \text{logit}(\theta_k) &= x'_k \beta. \end{aligned}$$

In a slight change of notation, write the predictor variables as two vectors  $x_k$  and  $z_k$  of dimensions  $r$  and  $q$ , respectively. If we similarly write the regression coefficients as two vectors  $\beta$  and  $\gamma$ , we can write a binomial regression model as

$$\begin{aligned} y_k | \theta_k &\stackrel{ind}{\sim} \text{Bin}(N_k, \theta_k) \\ \text{logit}(\theta_k) &= x'_k \beta + z'_k \gamma. \end{aligned}$$

The point is simply to have the regression coefficients, and their associated predictor variables, separated into two parts.

For a mixed model, we change the assumptions on one part of the regression coefficients. In standard regression, the coefficients are fixed unknown parameters. In a mixed model, some of the regression coefficients are assumed to be random. In particular, we assume for models with a fixed intercept

$$\gamma | \xi \sim N_q(0, \Sigma_\gamma(\xi))$$

with  $\Sigma_\gamma(\xi)$  a known positive definite matrix given the value of some parameter vector  $\xi$ . If the model  $x'_k \beta$  does not include an intercept, the elements of  $\gamma$  would have a common mean  $\mu$ .

In our most frequent applications, the random effects are due to the observations falling into groups, so the columns of  $z'_k$  consist of indicator variables for the groups (like hospitals). Moreover, the random group effects are viewed as exchangeable. For the various groups  $i$  and observations within groups  $j$ , consider a model

$$y_{ij} | \theta_{ij} \stackrel{ind}{\sim} \text{Bin}(N_i, \theta_{ij})$$

$$\begin{aligned} \text{logit}(\theta_{ij}) &= x'_{ij}\beta + \gamma_i \\ \gamma_i | \tau &\stackrel{iid}{\sim} N(0, 1/\tau). \end{aligned} \quad (2)$$

Again, if the  $x'_{ij}\beta$  model does not have an intercept, the elements of  $\gamma$  would have a common mean  $\mu$ . Let  $i = 1, \dots, a$  and  $j = 1, \dots, n_i$ .

**EXAMPLE 8.5.1. Toenail Fungus.** Toenail onychomycosis, known as toenail fungus, is fairly common. It can disfigure and sometimes destroy the nail. It may afflict between 2% and 18% of people world-wide. Onychomycosis can be caused by several types of fungi known as dermatophytes, as well as by non-dermatophytic yeasts or molds. Following Christensen et al. (2008) we consider data from a clinical trial on toenail fungus reported by De Backer et al. (1996). The randomized study compares two oral treatments for dermatophyte onychomycosis infection: terbinafine and intraconazole. These are well-known commercially available treatments.

Specifically, we consider data on an unpleasant side effect: the degree of separation of the nail plate from the nail bed. This is scored in four categories (0, absent; 1, mild; 2, moderate; 3, severe). For the  $a = 294$  patients, the response was evaluated at seven visits (approximately on weeks 0, 4, 8, 12, 24, 36, and 48). A total of 937 measurements were made on the 146 intraconazole,  $\text{Trt}_i = 0$ , patients and 971 measurements were made on 148 terbinafine,  $\text{Trt}_i = 1$ , patients. The data are available through the beneficence of Novoartis, Belgium at

<http://www.blackwellpublishing.com/rss/Volumes/Cv50p3.htm>

Alternatively, our website contains a link to the data as well as instructions on loading the data into WinBUGS.

One typical approach to analyzing ordinal data is to fit a series of logistic models such as continuation ratios or cumulative logits, see Christensen (1997, Section 4.6). Following Lesaffre and Spiessens (2001), we examine a logit mixed effects model. Specifically, let  $y_{ij} = 1$  if individual  $i$  has moderate or severe toenail separation at time  $j$ , with  $y_{ij} = 0$  if toenail separation is absent or mild. Consider the sampling model,

$$\begin{aligned} y_{ij} | \theta_{ij} &\stackrel{ind}{\sim} \text{Bern}(\theta_{ij}), \\ \text{logit}(\theta_{ij}) &= \gamma_i + \beta_1 \text{Trt}_i + \beta_2 \text{Time}_{ij} + \beta_3 \text{Trt}_i \times \text{Time}_{ij}, \\ \gamma_i | \mu, \tau &\stackrel{iid}{\sim} N(\mu, 1/\tau), \end{aligned}$$

for  $i = 1, \dots, 294$  and  $j = 1, \dots, n_i \leq 7$ . Here  $\gamma_i$  is a random effect for each subject.  $\text{Trt}$  is the binary treatment indicator.  $\text{Time}$  is the visit time as measured in four-week periods. (We treat  $\text{Time}$  as a continuous covariate although it could also be considered a factor with 7 levels.) The model also includes an interaction,  $\text{Trt} \times \text{Time}$ . The  $\beta_k$ s are fixed effects and do not include an intercept, so the random effects  $\gamma_i$  have a non-zero mean  $\mu$ . The  $\gamma_i$ s are included primarily to introduce correlation among the repeated observations on the same individual. Alternatively, each individual may have a biological tendency to have higher (or lower) probabilities of moderate to severe toenail separation through time.

To look at a typical individual, consider a case with  $\gamma = \mu$ . The log-odds are then straight line functions of  $\text{Time}$ , one line for each treatment, i.e.,

$$\text{logit}(\theta) = \begin{cases} \mu + \beta_2 \text{Time} & \text{if } \text{Trt} = 0 \\ (\mu + \beta_1) + (\beta_2 + \beta_3) \text{Time}, & \text{if } \text{Trt} = 1 \end{cases}.$$

Thus,  $\beta_1$  is the differential effect between the treatments at  $\text{Time} = 0$ . Since this is a randomized experiment, we have good reason to believe that  $\beta_1$  should be 0. Presuming that these treatments actually work, we should see the log-odds decreasing with time, thus we expect  $\beta_2$  and  $\beta_2 + \beta_3$  both to be negative. The effect  $\beta_3$  is the difference in slopes between the two treatments. A negative differential slope indicates that treatment 1 is better because the log-odds are decreasing faster with time.

Table 8.7: Posterior summaries for the toenail data.

Parameter	$\tau \sim \text{Gamma}(2, 0.5)$				
	mean	sd	.025	median	.975
$\beta_1(\text{Trt})$	-0.1715	0.5791	-1.337	-0.1597	0.9352
$\beta_2(\text{Time})$	-0.3926	0.04499	-0.4847	-0.3913	-0.3089
$\beta_3(\text{Trt} \times \text{Time})$	-0.1396	0.06805	-0.2757	-0.1388	-0.008348
$\mu$	-1.605	0.43	-2.493	-1.592	-0.7861
$\tau$	0.0642	0.0123	0.04289	0.06331	0.0915
Parameter	$\tau \sim \text{Gamma}(1.025, 0.01)$				
	mean	sd	.025	median	.975
$\beta_1(\text{Trt})$	-0.1226	0.6109	-1.312	-0.1332	1.117
$\beta_2(\text{Time})$	-0.3947	0.04576	-0.4882	-0.3934	-0.3084
$\beta_3(\text{Trt} \times \text{Time})$	-0.1388	0.0697	-0.2776	-0.1387	-0.001743
$\mu$	-1.658	0.4481	-2.592	-1.64	-0.8313
$\tau$	0.06246	0.01208	0.04172	0.0615	0.08869

We require priors for both the regression parameters  $\beta = (\beta_1, \beta_2, \beta_3)'$  and the parameters  $\mu$  and  $\tau$  of the normal distribution for the  $\gamma_i$ s. We used independent normal and gamma reference priors. Specifically,  $N_3(0, 100 \times I_3)$ ,  $N(0, 100)$ , and  $\text{Gamma}(2, 0.5)$  priors for  $\beta$ ,  $\mu$ , and  $\tau$ , respectively. An alternative choice of  $\tau \sim \text{Gamma}(1.025, 0.01)$  gave similar results. Since the sample size is large, the choice of priors for  $\beta$  and  $\mu$  will have little effect; however, that is not always the case for  $\tau$ . Estimating  $\tau$ , even for large samples sizes, is difficult if the large sample size is due to large  $n_i$ s with a relatively small number of units  $a$ . Here  $a = 294$  is reasonable.

Posterior summaries are given in Table 8.7 for both priors on  $\tau$ . Comparing the interval estimates to 0, the model shows no substantial baseline effect  $\beta_1$  for treatments, as should be the case for a randomized experiment. The model shows a negative slope for treatment 0 ( $\beta_2 < 0$ ) and that treatment 1 works even better over time ( $\beta_3 < 0$ ). The first prior gives posterior probability  $\Pr(\beta_3 < 0 | y) = 0.98$ .

The effectiveness of these treatments may bring some relief to the mushrooming fungus pandemic in Europe. Many families have been forced to seek relief in the south of France. Fortunately, my niece found a nice niche in Nice.

EXERCISE 8.20. (a) Treating Time as a factor, modify the WinBUGS code below to obtain estimates for the probabilities of moderate-severe separation for “typical” people in each of the 14 Time-Treatment categories. Construct a plot of the estimated probabilities over time for each of the treatment groups. (b) Perform a sensitivity analysis by perturbing the priors on  $\beta$ ,  $\mu$ , and  $\tau$ . (c) Modify the WinBUGS code below to include an intercept in the logistic regression and random effects with mean 0. Compare the results of the two methods including convergence properties.

```

model{
  for(i in 1:1908){
    y[i] ~ dbern(p[i])
    logit(p[i]) <- gamma[ID[i]] + beta[1]*Trt[i] + beta[2]*Time[i]
    + beta[3]*Trt[i]*Time[i]
  }
  # Note: ID[] identifies individuals,
  # 294 people listed as numbers between 1 and 383.
  for(i in 1:383){ gamma[i] ~ dnorm(mu,tau) }
  beta[1] ~ dnorm(0,0.01)
  beta[2] ~ dnorm(0,0.01)
  beta[3] ~ dnorm(0,0.01)
}

```

```

mu ~ dnorm(0,0.01)
tau ~ dgamma(2,0.5)
junk <- Visit[1]
}

```

### 8.5.1 Prior Elicitation

Consider a simple binomial regression for predicting hospital satisfaction from patient income  $x$  with random effect  $\gamma$ , say,  $\text{logit}(\theta) \equiv \mu + \beta x + \gamma$ , where  $\gamma \sim N(0, \sigma^2)$  and  $\tau = 1/\sigma^2$ . As usual, the prior on  $\mu$  and  $\beta$  is obtained by eliciting information at two values  $\tilde{x}_1$  and  $\tilde{x}_2$ , but now this is elicited for a “typical” individual, i.e., one with  $\gamma = 0$ . The extension to multiple binomial regression is straightforward. The remainder of the discussion does not really depend on the number of predictors.

To elicit information on  $\sigma^2$ , we start by fixing  $x$  at  $x_*$ . This may be  $\tilde{x}_1$  or  $\tilde{x}_2$  but not necessarily. Obtain a best guess for the corresponding  $\theta$ , say  $\theta_*$ . Our best guess for  $\mu + \beta x_*$  is then  $\text{logit}(\theta_*)$ .

Now think about satisfaction probabilities across all hospitals with covariate  $x_*$ . (Thus, we are no longer conditioning on  $\gamma$ .) We ask our expert for their best guess of, say, the 90th percentile of probabilities for such hospitals. What does that mean? Imagine we have 100 hospitals with covariate  $x_*$ . Then we expect about 90 of these 100 hospitals to have satisfaction probabilities below, say,  $\theta_{0.9}$ . Suppose our expert’s best guess is  $\theta_{0.9*} = 0.7$ . Then, we think that 90% of hospitals with  $x_*$  will have satisfaction probabilities below 0.7. From the normality of  $\gamma$ , we must have  $\text{logit}(\theta_{0.9}) = \mu + \beta x_* + 1.28\sigma$ . Since our best guess for  $\mu + \beta x_*$  is  $\text{logit}(\theta_*)$ , our best guess for  $\text{logit}(\theta_{0.9})$  must satisfy  $\text{logit}(\theta_{0.9*}) = \text{logit}(\theta_*) + 1.28\sigma$ . Solving for  $\sigma$ , we get a best guess for  $\sigma$  of  $\sigma_* = [\text{logit}(\theta_{0.9*}) - \text{logit}(\theta_*)]/1.28$  or equivalently, a best guess for  $\tau$  of  $\tau_* = \{1.28/[\text{logit}(\theta_{0.9*}) - \text{logit}(\theta_*)]\}^2$ .

Placing a  $\text{Gamma}(a, b)$  prior on either  $\sigma$  or  $\tau$ , we can set  $\sigma_* = (a-1)/b$  or  $\tau_* = (a-1)/b$ , and solve for  $a$  in terms of  $b$  as we did in Chapter 5. We can then be lazy and let  $b$  be a small number corresponding to a large prior variance. We would generally not pick  $b$  to be too small since otherwise we might place too much prior probability on very large  $\sigma$  or small  $\tau$ , in which case  $\gamma$  could be large or small with high probability and thus induce probabilities close to 0 and 1 on probabilities  $\theta$ . It would be wise to try a variety of values for  $b$  and look at the corresponding priors on the  $\theta_i$  of the data.

Alternatively, we can continue eliciting information about the 90th percentile of  $\theta$ . Eliciting information that leads to a choice of  $b$  involves eliciting a value that the expert is, say, 95% sure that  $\theta_{0.9}$  could not exceed, say  $\theta_{0.9u}$ , or a corresponding lower value, say  $\theta_{0.9l}$ . In the former case, we get

$$\begin{aligned}
0.95 &= \Pr[\theta_{0.9} \leq \theta_{0.9u} | \text{logit}(\theta_*)] \\
&= \Pr[\mu + x_*\beta + 1.28\sigma \leq \text{logit}(\theta_{0.9u}) | \text{logit}(\theta_*)] \\
&= \Pr[\text{logit}(\theta_*) + 1.28\sigma \leq \text{logit}(\theta_{0.9u})] \\
&= \Pr[\sigma \leq (\text{logit}(\theta_{0.9u}) - \text{logit}(\theta_*))/1.28] \\
&= \Pr[\tau \geq \{1.28/(\text{logit}(\theta_{0.9u}) - \text{logit}(\theta_*))\}^2]
\end{aligned}$$

where we have assumed that  $\mu + \beta x_*$  is independent of  $\sigma$ . In the case of a  $\text{Gamma}(a, b)$  prior on  $\sigma$ , we now find  $b$  so that the  $\text{Gamma}(1 + \sigma_* b, b)$  distribution has 95th percentile equal to  $(\text{logit}(\theta_{0.9u}) - \text{logit}(\theta_*))/1.28$ . Similarly, we could use the 5th percentile for  $\tau$  or elicit the 5th and 95th percentiles for  $\sigma$  and  $\tau$ , respectively, by working with  $\theta_{0.9l}$ .

A third alternative is to place a uniform prior on  $\sigma$  by finding an upper limit for  $\theta_{0.9}$  that we believe is impossible to exceed, say  $\theta_u$ . Using exactly the same logic as above, we find that  $\theta_{0.9} \leq \theta_u$  implies

$$\sigma \leq \frac{\text{logit}(\theta_u) - \text{logit}(\theta_*)}{1.28} \equiv u$$

and let  $\sigma \sim U(0, u)$ .

EXERCISE 8.21. The following code handles a mixed logistic regression model where there are 0/1 responses on individual cows coming from nine herds located in the central valley of California. The data were collected by Drs. Mark Thurmond and Sharon Hietala of the Veterinary School at the University of California, Davis. Prior information was elicited from Dr. Thurmond. Let  $y = 1$  correspond to “natural abortion” and  $y = 0$  correspond to no abortion. There are two covariate values for each of the 13,145 cows. These are  $GR$  = gravidity (the number of previous successful pregnancies) and  $DO$  = days open (the number of days between the last successful birth and the current pregnancy). We expect that higher  $DO$ s will correlate with higher probabilities of natural abortion since a longer  $DO$  may be a result of a difficult calving for the previous birth; and that higher  $GR$  will result in lower probability of natural abortion. We assume a random herd effect. The goal is to model the probability of abortion as a function of the covariates and a random effect for herd. We expect that the effect of  $DO$  may be modified by  $GR$ . The data are available on our website.

(a) The code given below uses a slightly informative prior. The model presented there corresponds to centering the distribution of random effects on  $\mu$ , and has correspondingly left out the intercept. The prior on  $\mu$  reflects a belief that the average prevalence of abortion among cows is about 12%. Dr. Thurmond’s best guess for  $\sigma$  is about 0.25, based on the reasoning used in this section. Relative to that belief, the prior used is a very diffuse  $U(0, 2)$  distribution.

Dr. Thurmond wants to know the effect of the covariates on the probability of abortion over the range of covariate values in the data, for typical herds, e.g., with random effects set to  $\mu$ . He is also interested in knowing the effect of herds on these probabilities. That is because certain practices in maintaining herds, not accounted for by  $GR$  or  $DO$ , may result in either higher (with poor practices) or lower (with good practices) probabilities of abortion. Run the code on the full data. You may wish to augment the code to obtain additional inferences. *Be warned that the code takes about four minutes per thousand iterations.*

(b) Modify the code to the standard parameterization where the random effects are centered at zero, and then include the intercept. Run the code for 1,000 iterations or so. What do you notice? The Markov chains for all variables except the random effects will have the same stationary distribution. Centering as in part (a) often results in better behaved chains. See Subsection 8.5.3 for additional details.

(c) With Dr. Thurmond’s prior information given in (a), explain why the coded normal distribution is appropriate. Give an appropriate Gamma distribution for  $\sigma$  that incorporates the prior information, is still diffuse, but not overly so. Modify the code to use the same prior for  $\mu$ , the new prior for  $\sigma$ , and use `dflat()` for the regression coefficients. Run the modified code and explain how inferences would differ with this prior compared to the one in part (a).

(d) (i) Find a prior for  $\sigma$  that corresponds to  $\theta_* = 0.12$  (evaluated at average  $GR$  and average  $DO$ ),  $\theta_{0.9*} = 0.20$ , and  $\theta_{0.9u} = 0.24$ . (ii) Find a uniform prior for  $\sigma$  that corresponds to  $\theta_u = 0.3$ .

```
model{
  meangr <- mean(GR[ ])
  meando <- mean(DO[ ])
  stdgr <- sd(GR[ ])
  stddo <- sd(DO[ ])
  for(k in 1:13145){
    grstd[k] <- (GR[k] - meangr)/stdgr
    dostd[k] <- (DO[k] - meando)/stddo
  }
  for(k in 1:9) { gamma[k] ~ dnorm(mu, tau) }
  tau <- 1/pow(sigma,2)
  sigma ~ dunif(0,10)
  for(k in 1:13145){
    y[k] ~ dbern(pr[k])
    logit(pr[k]) <- beta[1]*dostd[k] + beta[2]*grstd[k]
```

```

+ beta[3]*dostd[k]*grstd[k] + gamma[herd[k]]
}
mu ~ dnorm (-2,0.001)
for(i in 1:3){ beta[i] ~ dnorm(0,0.01) }
}
list(beta=c(-0.000867,-0.2093,-0.000215), sigma=1, mu=-2)
herd[ ] DO[ ] GR[ ] y[ ]
1      31      3      0
1      31      5      1
1      31      3      0
remaining data lines go here
END

```

### 8.5.2 Mixed Model Likelihood

The likelihood function of a general binomial mixed model is not too different from that of the model defined by (2). In this special case, the likelihood is computed directly as

$$L(\beta, \mu, \tau) = \prod_i \int \prod_j \left( \frac{e^{x'_{ij}\beta + \gamma_i}}{1 + e^{x'_{ij}\beta + \gamma_i}} \right)^{y_{ij}} \left( \frac{1}{1 + e^{x'_{ij}\beta + \gamma_i}} \right)^{1-y_{ij}} \tau^{\frac{1}{2}} e^{-\tau(\gamma_i - \mu)^2/2} d\gamma_i.$$

The integral over  $\gamma_i$  is the likelihood contribution for the  $i$ th group, e.g., the  $i$ th hospital or individual.

Alternatively, we can treat the  $\gamma_i$ s as parameters. Putting them into a vector  $\gamma$  leads to

$$L(\beta, \gamma, \mu, \tau) = \prod_i \prod_j \left( \frac{e^{x'_{ij}\beta + \gamma_i}}{1 + e^{x'_{ij}\beta + \gamma_i}} \right)^{y_{ij}} \left( \frac{1}{1 + e^{x'_{ij}\beta + \gamma_i}} \right)^{1-y_{ij}}.$$

Then, as a step in a Bayesian analysis to find the posterior of  $\beta$ ,  $\mu$ , and  $\tau$ , we equivalently obtain

$$\begin{aligned} L(\beta, \mu, \tau) &= \int f(y|\beta, \gamma, \mu, \tau) p(\gamma|\beta, \mu, \tau) d\gamma = \int L(\beta, \gamma, \mu, \tau) p(\gamma|\beta, \mu, \tau) d\gamma \\ &= \int \prod_i \prod_j \left( \frac{e^{x'_{ij}\beta + \gamma_i}}{1 + e^{x'_{ij}\beta + \gamma_i}} \right)^{y_{ij}} \left( \frac{1}{1 + e^{x'_{ij}\beta + \gamma_i}} \right)^{1-y_{ij}} \left[ \prod_i \tau^{\frac{1}{2}} e^{-\tau(\gamma_i - \mu)^2/2} d\gamma_i \right] \\ &= \prod_i \int \prod_j \left( \frac{e^{x'_{ij}\beta + \gamma_i}}{1 + e^{x'_{ij}\beta + \gamma_i}} \right)^{y_{ij}} \left( \frac{1}{1 + e^{x'_{ij}\beta + \gamma_i}} \right)^{1-y_{ij}} \tau^{\frac{1}{2}} e^{-\tau(\gamma_i - \mu)^2/2} d\gamma_i. \end{aligned}$$

The integrand of the penultimate expression is the *augmented data likelihood*. This times the prior  $p(\beta, \mu, \tau)$  is used to obtain full conditional distributions, cf. Subsection 8.5.3.

### 8.5.3 Gibbs Sampling and Centering\*

Complicated models, such as binomial mixed models, typically have many alternative parameterizations. As discussed in Section 4.13, it is not clear whether  $\gamma$  should be treated as a parameter vector or whether it should be integrated out prior to the analysis. Effective sampling from the posterior often depends on the exact parameterization. We now illustrate how two parameterizations for binomial mixed models affect the Gibbs sampler. Both treat  $\gamma$  as a parameter rather than integrating it out.

Suppose

$$y_1, \dots, y_n | \theta_i \stackrel{ind}{\sim} \text{Bin}(N_i, \theta_i), \quad \text{logit}(\theta_i) \equiv x'_i \beta + \gamma_i,$$

and

$$\gamma_1, \dots, \gamma_n | \tau \stackrel{iid}{\sim} N(0, 1/\tau).$$

This is a special case of the model in (2) with  $j \equiv 1$ .

Although we do not typically recommend normal priors for  $\beta$ , for illustration take

$$\beta \sim N(\beta_0, \Sigma_0) \quad \perp \quad \tau \sim \text{Gamma}\left(\frac{a}{2}, \frac{b}{2}\right).$$

The joint density has the form

$$p(y, \beta, \gamma, \tau) = f_*(y|\beta, \gamma) f_0(\gamma|\tau) p(\beta) p(\tau).$$

Specifically, the joint density is

$$\begin{aligned} & p(y, \beta, \gamma, \tau) \\ & \propto \left[ \prod_{i=1}^n \left( \frac{e^{x_i' \beta + \gamma_i}}{1 + e^{x_i' \beta + \gamma_i}} \right)^{y_i} \left( \frac{1}{1 + e^{x_i' \beta + \gamma_i}} \right)^{N_i - y_i} \right] \\ & \quad \times \left[ \prod_{i=1}^n \tau^{1/2} e^{-\tau \gamma_i^2 / 2} \right] \\ & \quad \times \left[ \exp \left( -\frac{1}{2} (\beta - \beta_0)' \Sigma_0^{-1} (\beta - \beta_0) \right) \right] \\ & \quad \times \left[ \tau^{\frac{a}{2}-1} e^{-b\tau/2} \right]. \end{aligned}$$

To perform Gibbs sampling we need to be able to find

$$p(\beta|y, \gamma, \tau), \quad p(\gamma|y, \beta, \tau), \quad p(\tau|y, \beta, \gamma).$$

From the form of the joint density,

$$p(\tau|y, \beta, \gamma) \sim \text{Gamma}\left(\frac{n+a}{2}, \frac{b+\sum \gamma_i^2}{2}\right).$$

We can also see that the  $\gamma$ s are conditionally independent with

$$p(\gamma|y, \beta, \tau) \propto \left( \frac{e^{x_i' \beta + \gamma_i}}{1 + e^{x_i' \beta + \gamma_i}} \right)^{y_i} \left( \frac{1}{1 + e^{x_i' \beta + \gamma_i}} \right)^{N_i - y_i} \tau^{\frac{1}{2}} e^{-\tau \gamma_i^2 / 2}.$$

This is not a recognizable distribution but at least it is a scalar distribution, making it easier to sample. Finally, we could use Metropolis or Acceptance-Rejection to sample from the multivariate distribution

$$\begin{aligned} p(\beta|y, \gamma, \tau) & \propto \left[ \prod_{i=1}^n \left( \frac{e^{x_i' \beta + \gamma_i}}{1 + e^{x_i' \beta + \gamma_i}} \right)^{y_i} \left( \frac{1}{1 + e^{x_i' \beta + \gamma_i}} \right)^{N_i - y_i} \right] \\ & \quad \times \left[ \exp \left( -\frac{1}{2} (\beta - \beta_0)' \Sigma_0^{-1} (\beta - \beta_0) \right) \right]. \end{aligned}$$

Alternatively, we could sample each scalar  $\beta_j$  by sampling  $p(\beta_j|y, \beta_{-j}, \gamma, \tau)$  for  $j = 1, \dots, r$  where  $\beta_{-j}$  denotes the  $r-1$  vector that is  $\beta$  with  $\beta_j$  removed.

Because there is a separate random effect for each individual, we can modify the parameterization by putting the fixed regression effects into the model for the random effects. This redefining of

the  $\gamma_i$ s is called the *centering model*. It does not affect the substance of the model, the likelihood function is identical, but it affects how we sample from the posterior.

Take

$$y_i | \gamma \stackrel{\text{ind}}{\sim} \text{Bin}(N_i, \theta_i), \quad \text{logit}(\theta_i) \equiv \gamma_i,$$

and

$$\gamma_1, \dots, \gamma_n | \beta, \tau \stackrel{\text{ind}}{\sim} N(x_i' \beta, 1/\tau). \quad (3)$$

Now  $\beta$  is a parameter relative to  $\gamma$  rather than  $y_i$ . We can only do this because the model includes a separate parameter  $\gamma_i$  for each  $y_i$ , something that does not generally happen in mixed models for binomial data. Moreover (3) is a linear model and amenable to exact analysis as discussed in Subsection 9.4.5.

Using the same prior

$$\beta \sim N(\beta_0, \Sigma_0) \quad \perp\!\!\!\perp \quad \tau \sim \text{Gamma}\left(\frac{a}{2}, \frac{b}{2}\right),$$

the joint density has the form

$$p(y, \beta, \gamma, \tau) = f_*(y|\gamma) f_0(\gamma|\beta, \tau) p(\beta) p(\tau).$$

Specifically, the joint density is

$$\begin{aligned} p(y, \beta, \gamma, \tau) & \\ &\propto \left[ \prod_{i=1}^n \left( \frac{e^{\gamma_i}}{1+e^{\gamma_i}} \right)^{y_i} \left( \frac{1}{1+e^{\gamma_i}} \right)^{N_i-y_i} \right] \\ &\quad \times \left[ \prod_{i=1}^n \tau^{1/2} e^{-\tau(\gamma_i - x_i' \beta)^2/2} \right] \\ &\quad \times \left[ \exp\left(-\frac{1}{2}(\beta - \beta_0)' \Sigma_0^{-1} (\beta - \beta_0)\right) \right] \\ &\quad \times \left[ \tau^{a/2-1} e^{-b\tau/2} \right]. \end{aligned}$$

To perform Gibbs sampling we need to find

$$p(\beta|y, \gamma, \tau), \quad p(\gamma|y, \beta, \tau), \quad p(\tau|y, \beta, \gamma).$$

The full conditional for  $\tau$  is

$$p(\tau|y, \beta, \gamma) \sim \text{Gamma}\left(\frac{n+a}{2}, \frac{b + \sum(\gamma_i - x_i' \beta)^2}{2}\right).$$

Again we can see that the  $\gamma_i$ s are conditionally independent with

$$p(\gamma_i|y, \beta, \tau) \propto \left( \frac{e^{\gamma_i}}{1+e^{\gamma_i}} \right)^{y_i} \left( \frac{1}{1+e^{\gamma_i}} \right)^{N_i-y_i} \tau^{1/2} e^{-\tau(\gamma_i - x_i' \beta)^2/2}.$$

Again, this is not recognizable, but is a scalar distribution.

Mathematically, the big advantage is that the full conditional distribution of  $\beta$  is a multivariate normal distribution. The full conditional is based on

$$p(\beta|y, \gamma, \tau) \propto \left[ \prod_{i=1}^n \tau^{1/2} e^{-\tau(\gamma_i - x_i' \beta)^2/2} \right] \left[ \exp\left(-\frac{1}{2}(\beta - \beta_0)' \Sigma_0^{-1} (\beta - \beta_0)\right) \right].$$

This does not depend on the  $y_i$ s; it is actually a standard normal theory linear model posterior as discussed in Chapter 9 where  $\tau$  is known and the data are the  $\gamma$ s. With the matrix  $X$  having rows  $x'_i$ , from results derived in Chapter 9 the posterior multivariate normal has mean vector

$$\tilde{\beta}(\gamma, \tau) = (\tau X'X + \Sigma_0^{-1})^{-1} [\tau X'\gamma + \Sigma_0^{-1}\beta_0]$$

and covariance matrix  $(\tau X'X + \Sigma_0^{-1})^{-1}$ . Of course this only occurs because we are using a normal prior for  $\beta$ , a prior that we do not recommend.

The centering model is used to alleviate correlations among MCMC iterates and improve convergence in the chain. There is no guarantee that centering helps but in our experience it does.

More generally, in model (2) we would partition  $x_{ij}$  into  $x'_{ij} = (x'_{1ij}, x'_{2i})$  where the second component includes only “group” variables (including the intercept) that do not change with  $j$ . Using a conformable partition  $\beta' = (\beta'_1, \beta'_2)$ , the centering model becomes

$$y_{ij} | \theta_{ij} \stackrel{ind}{\sim} \text{Bin}(N_i, \theta_{ij}), \quad \text{logit}(\theta_{ij}) \equiv x'_{1ij}\beta_1 + \gamma_i,$$

and

$$\gamma_1, \dots, \gamma_n | \beta, \tau \stackrel{ind}{\sim} N(x'_{2i}\beta_2, 1/\tau).$$

With a normal prior on  $\beta_2$ , we can sample  $\beta_2$  from a multivariate normal while using, say, Metropolis for  $\beta_1$ .

---

## Chapter 9

---

# Linear Regression

---

Linear regression is about predicting a continuous response variable  $y$  from one or more predictor variables  $x$ . The linear regression model specifies the mean of the response variable as a linear combination of the predictor variables:  $E[y|x] = x'\beta$ . Although  $x$  is often random, we *always* condition on it in this chapter, hence we treat it as fixed. Since covariate combinations can consist of continuous variables or categorical variables, standard ANOVA and ACOVA models are special cases of these linear models. Theoretical results in this chapter rely heavily on multivariate normality, cf. Example B.1.

Section 9.1 discusses the sampling distribution of the data. Sections 9.2–9.4 present several types of priors, including procedures for eliciting informative priors, along with inferential reasoning and computational implementation in WinBUGS and R. ANOVA is discussed in Section 9.5. Model diagnostics and covariate selection are detailed in Sections 9.6 and 9.7, respectively, and nonlinear regression is introduced in Section 9.8.

Standard notation for the matrix representation of linear models will be used in Chapters 9 and 10. In particular,  $Y$  will denote the vector of responses rather than our usual notation  $y$ .

### 9.1 The Sampling Model

We begin with an example.

**EXAMPLE 9.1.1. Bank Salaries.** Schafer (1987) reported data from 1977 on 93 bank employees. The response variable is beginning salary in dollars. Along with the intercept, there are four covariates:  $x_2$  – sex, i.e., an indicator variable for “male”;  $x_3$  – years of education;  $x_4$  – months of experience;  $x_5$  – time at hiring as measured in months after January 1, 1969. Figure 9.1 gives a scatterplot matrix of the variables other than sex. These separate bivariate plots indicate that beginning salary is positively correlated with education, months of experience, and time of hiring.

The typical linear model specifies that for known predictor variables  $x_i' = (1, x_{i2}, \dots, x_{ir})$  and unknown regression coefficients  $\beta = (\beta_1, \beta_2, \dots, \beta_r)'$ ,

$$\begin{aligned} y_i &= \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \dots + \beta_r x_{ir} + \varepsilon_i \\ &= x_i' \beta + \varepsilon_i \\ \varepsilon_i | \tau &\stackrel{iid}{\sim} N(0, 1/\tau), \end{aligned} \tag{1}$$

for  $i = 1, \dots, n$ . Equivalently,

$$y_i | \beta, \tau \stackrel{ind}{\sim} N(x_i' \beta, 1/\tau). \tag{2}$$

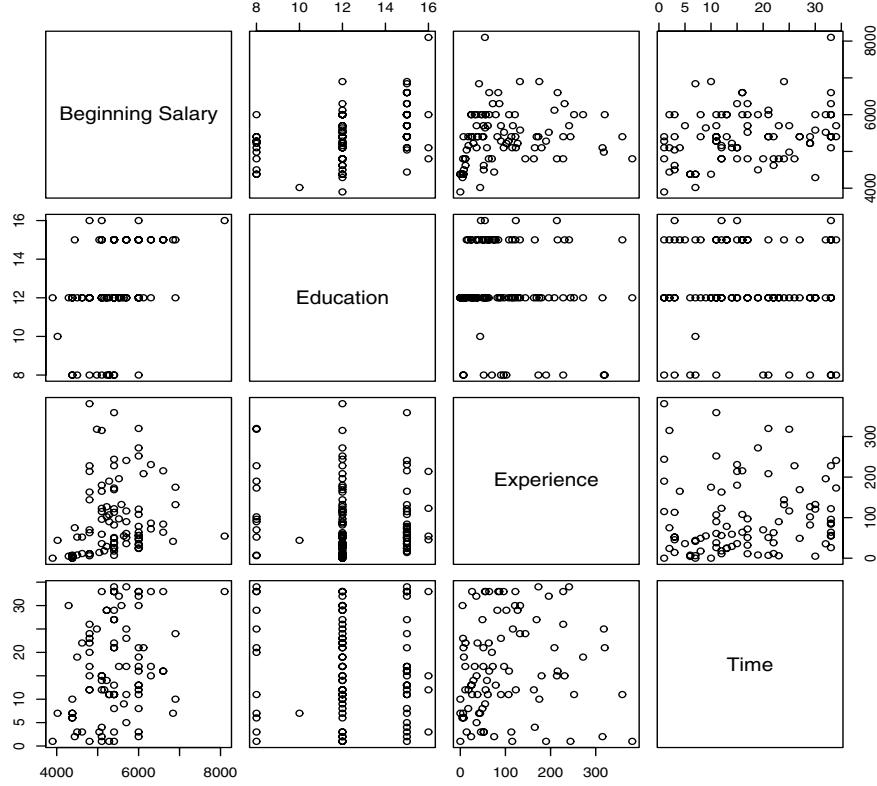


Figure 9.1: Scatterplot matrix for bank salary data.

The model can be written completely in matrix form as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{12} & x_{13} & \cdots & x_{1r} \\ 1 & x_{22} & x_{23} & \cdots & x_{2r} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n2} & x_{n3} & \cdots & x_{nr} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_r \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$Y_{n \times 1} = X_{n \times r} \beta_{r \times 1} + \varepsilon_{n \times 1}$$

It is common for the first column of  $X$  to consist of 1s in order to accommodate an intercept parameter but there is no requirement for it. Computing the right-hand side of the matrix equation gives

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \beta_1 + \beta_2 x_{12} + \beta_3 x_{13} + \cdots + \beta_r x_{1r} + \varepsilon_1 \\ \beta_1 + \beta_2 x_{22} + \beta_3 x_{23} + \cdots + \beta_r x_{2r} + \varepsilon_2 \\ \vdots \\ \beta_1 + \beta_2 x_{n2} + \beta_3 x_{n3} + \cdots + \beta_r x_{nr} + \varepsilon_n \end{bmatrix},$$

with rows determined by equation (1). The conditions on the  $\varepsilon_i$ s imply that

$$E(\varepsilon) = 0,$$

where 0 is the  $n \times 1$  matrix consisting of all zeros, and

$$\text{Cov}(\varepsilon) = \tau^{-1} I_n,$$

where  $I_n$  is the  $n \times n$  identity matrix. In particular, given  $\tau, \varepsilon$  has the multivariate normal distribution of Example B.1,

$$\varepsilon | \tau \sim N_n(0, \tau^{-1} I_n).$$

Succinctly, the multiple linear regression model is

$$Y = X\beta + \varepsilon, \quad \varepsilon | \tau \sim N_n(0, \tau^{-1} I_n).$$

What makes this specifically a regression model (rather than an arbitrary linear model) is that we require  $X$  to have full column rank, i.e.,  $\text{rank}(X) = r$ . Unless  $n \geq r$ , the problem degenerates because there are more parameters than observations and  $\text{rank}(X) < r$ . As in Section B.2, it follows from simple rules about expected values and covariances of linear transformations that  $E(Y) = X\beta$ ,  $\text{Cov}(Y) = \tau^{-1} I_n$ , and therefore the linear regression sampling model is also written

$$Y | \beta, \tau \sim N_n(X\beta, \tau^{-1} I_n).$$

From (2), the likelihood as a function of  $\beta$  and  $\tau$  is given by

$$\begin{aligned} L(\beta, \tau) &\propto \prod_{i=1}^n \tau^{1/2} \exp \left\{ -\frac{\tau}{2} (y_i - x_i' \beta)^2 \right\} \\ &= \tau^{n/2} \exp \left\{ -\frac{\tau}{2} \sum_{i=1}^n (y_i - x_i' \beta)^2 \right\} \\ &= \tau^{n/2} \exp \left\{ -\frac{\tau}{2} (Y - X\beta)' (Y - X\beta) \right\}. \end{aligned} \quad (3)$$

The last equality follows because  $(Y - X\beta)' (Y - X\beta)$  is the sum of the squares of the elements in the vector

$$(Y - X\beta) = \begin{bmatrix} y_1 - x_1' \beta \\ \vdots \\ y_n - x_n' \beta \end{bmatrix}.$$

In Subsection 9.2.1 we establish that the maximum likelihood estimate (MLE) of  $\beta$  is

$$\hat{\beta} = (X'X)^{-1} X' Y$$

and that the MLE of  $\tau$  is  $\hat{\tau} = n/SSE$  where

$$SSE \equiv (Y - X\hat{\beta})' (Y - X\hat{\beta})$$

is the *sum of squares for error*. The standard unbiased (frequentist) estimator of  $\sigma^2 = 1/\tau$  is the *mean squared error (MSE)* defined by

$$MSE \equiv SSE/dfe \quad \text{where} \quad dfe \equiv (n - r)$$

is the *degrees of freedom for error*.

Simple linear regression predicts a continuous (measurement) response using only an intercept and one predictor variable, so for simple linear regression the likelihood is a function of the 3 parameters  $\beta_1, \beta_2, \tau$ . Multiple linear regression simply has more than one predictor variable in which case the likelihood is a function of the  $r+1$  parameters  $\theta = (\beta_1, \dots, \beta_r, \tau)'$ .

**EXERCISE 9.1.** *Simple Linear Regression.* Write the simple linear regression model using vector and matrix notation. Obtain an explicit formula for  $\hat{\beta}$ .

We need to specify a prior for  $\theta' = (\beta', \tau)$ . We discuss five types of prior specifications: (i) the *standard improper reference (SIR) prior*, (ii) a diffuse proper reference prior, (iii) the conjugate normal-gamma prior, (iv) an informative BCJ prior with  $\beta$  and  $\tau$  independent, and (v) a partially informative prior.

## 9.2 Reference Priors

The standard reference prior for linear regression analysis is improper. The SIR prior is

$$p(\beta, \tau) = 1/\tau,$$

which does not integrate to 1. SIR is not to be confused with “Sampling Importance Resampling” or “Sliced Inverse Regression.” The frequentist regression estimates given by virtually all computer packages are least squares estimates. Least squares estimates are also the estimates that correspond to this prior. In the next subsection we define least squares estimates, summarize key distributional results, and prove the formula for the estimates. In the following subsection, we relate those results to Bayesian inference using the SIR prior. We begin by illustrating the analysis using the SIR prior.

**EXAMPLE 9.2.1.** *Bank Salaries Continued.* Posterior results using a SIR prior for a regression analysis of the salary data of bank employees in 1977 are given in Table 9.1. In addition to inferences for  $\beta$  and  $\tau$ , the table also includes predictive inference for a new observation  $y_f$  with covariate vector corresponding to a male with 16 years of school, 54.5 months of experience, and hired 33 months after the start date for the study, i.e.,  $x'_f = (1, 1, 16, 54.5, 33.0)$  with corresponding mean  $x'_f\beta$ . The vector  $x_f$  was chosen because it has the same covariates as the highest paid individual in the data, a man making \$8,100.

Table 9.1: Regression on bank salary data with a SIR prior.

Parameter	Mean	Scale	2.5%	97.5%
Constant	3526.4	327.7	2875.2	4177.6
Sex	722.3	117.8	488.2	956.4
Educ	90.02	24.69	40.95	139.09
Exp	1.2679	0.5871	0.10	2.43
Time	23.428	5.200	13.09	33.76
$x'_f\beta$	6531.2	145.7	6241.6	6820.8
$y_f$	6531.2	527.9	5482.0	7580.4
$\tau$	$3.88 \times 10^{-6}$	$0.585 \times 10^{-6}$	$2.82 \times 10^{-6}$	$5.11 \times 10^{-6}$

Except for  $\tau$ , all of the parameters have marginal posterior  $t(88)$  distributions and the posterior means and scales are exactly as in least squares output. For parameters with  $t$  distributions, rather than reporting the posterior standard deviation, we report the scale parameter of the  $t$  distribution. The standard deviation is  $\sqrt{dfe/(dfe - 2)}$  times the scale parameter. The percentiles are obtained as  $\text{Mean} \pm 1.987(\text{Scale})$  where  $1.987 = t(0.975, 88)$ . The squared sample correlation between the 93 actual observations and their predicted values is  $R^2 = 0.51 = 51\%$ , which means that 51% of the variability in the salaries is explained by this regression.

A superficial interpretation of the output suggests that we are 95% sure a male makes, on average, between \$488 and \$956 more than his female counterpart who has the same education, experience, and month of hiring. We are also 95% sure that each additional year of education corresponds to a higher starting salary, on average, of between \$41 to \$139 with other variables fixed, and so on. As tempting as it may be, *these are not to be interpreted as causal relationships*. There is no reason to believe that if someone increased their education by a year that their salary would be increased.

From the 95% prediction interval, the future individual will make a salary between \$5,482 and \$7,580. The person in the sample with these covariates might have been overpaid relative to these qualifications.

For inference about  $\sigma$ , we start with inference for  $\tau$ . Summary inferences for  $\tau$  are given in Table 9.1. We have  $dfe = 88$  and  $SSE = 22,657,464$ , so  $MSE = 257,471 = 22,657,464/88$  and

$$\hat{\tau} = E(\tau | Y) = 1/(257,471) = 3.88 \times 10^{-6}.$$

The posterior standard deviation of  $\tau$  is  $\hat{\tau}\sqrt{2/dfe}$ , so

$$0.585 \times 10^{-6} = (3.88 \times 10^{-6}) \sqrt{2/88}.$$

To find percentiles of the posterior distribution for  $\tau$ , divide the corresponding percentiles of a  $\chi^2_{88}$  by  $SSE$ . The chi-squared percentiles for a 95% PI are  $\chi^2_{88}(0.025) = 63.941$  and  $\chi^2_{88}(0.975) = 115.841$ , so the interval for  $\tau$  has endpoints

$$2.82 \times 10^{-6} = \frac{63.941}{22,657,464} \quad \text{and} \quad 5.11 \times 10^{-6} = \frac{115.841}{22,657,464}.$$

Finally, take the 95% PI for  $\tau$  and convert it into a 95% PI for  $\sigma$  by taking inverse square roots, i.e.,

$$10^3(\sqrt{1/5.11}, \sqrt{1/2.82}) = (442.37, 595.49).$$

The sampling model should always be validated using the standard array of regression diagnostics as discussed in Section 9.6.

### 9.2.1 Least Squares Estimation

There is an intimate relationship between Bayesian and frequentist inference for linear models when using the SIR prior. Many aspects are identical, modulo the difference in interpretation. Algebraic calculations used in frequentist analysis are also useful for developing the Bayesian approach, so we provide a brief version of the frequentist analysis first. See Christensen (2002) for details related to this subsection.

In simple linear regression the least squares estimates are defined to be the values of  $\beta_1$  and  $\beta_2$  that minimize

$$\sum_{i=1}^n (y_i - \beta_1 - \beta_2 x_i)^2.$$

For multiple regression, least squares estimates minimize

$$\sum_{i=1}^n (y_i - \beta_1 - \beta_2 x_{i2} - \beta_3 x_{i3} - \cdots - \beta_r x_{ir})^2$$

with respect to the  $\beta_j$ s, or equivalently minimize

$$\sum_{i=1}^n (y_i - x'_i \beta)^2 = (Y - X\beta)'(Y - X\beta).$$

For any fixed  $\tau$ , the likelihood in equation (9.1.3) is maximized by minimizing  $(Y - X\beta)'(Y - X\beta)$ . We will see that the least squares estimate,  $\hat{\beta}$ , does not depend on  $\tau$ , so it is also the MLE. Replacing  $\beta$  with  $\hat{\beta}$  in (9.1.3), it becomes a one-dimensional calculus problem to show that the MLE of  $\tau$  is  $n/SSE$ .

The sampling distribution of the least squares estimate is

$$\hat{\beta} | \beta, \sigma^2 \sim N(\beta, \sigma^2(X'X)^{-1}) \tag{1}$$

and it is also well known that

$$\frac{(Y - X\hat{\beta})' (Y - X\hat{\beta})}{\sigma^2} \Big| \sigma^2 \sim \chi^2_{n-r} = \text{Gamma}\left(\frac{n-r}{2}, \frac{1}{2}\right). \tag{2}$$

$MSE$  is an unbiased estimate of  $\sigma^2$ , i.e.,  $E[MSE | \beta, \sigma^2] = \sigma^2$ .

Many interesting functions of the parameter  $\beta$  can be written as  $c'\beta$  for some known vector  $c$ . For example,  $\beta_1$ ,  $\beta_2$ , and  $\beta_2 - \beta_3$  can be written this way. The key distributional result is

$$\frac{c'\hat{\beta} - c'\beta}{\sqrt{MSE c'(X'X)^{-1}c}} \Big| \beta \sim t(n-r). \quad (3)$$

When predicting a new random observation  $y_f$  with known covariate vector  $x_f$ , the relevant distribution is

$$\frac{y_f - x'_f \hat{\beta}}{\sqrt{MSE[1 + x'_f(X'X)^{-1}x_f]}} \sim t(n-r). \quad (4)$$

In all these distributions,  $Y$ , and thus  $\hat{\beta}$ ,  $SSE$ , and  $MSE$ , are considered random, while the parameters  $\beta$  and  $\sigma^2$  are considered fixed.

We now show that the least squares estimate of  $\beta$  is  $\hat{\beta} = (X'X)^{-1}X'Y$ . Rewrite the function to be minimized as

$$\begin{aligned} & (Y - X\beta)'(Y - X\beta) \\ &= (Y - X\hat{\beta} + X\hat{\beta} - X\beta)'(Y - X\hat{\beta} + X\hat{\beta} - X\beta) \\ &= (Y - X\hat{\beta})'(Y - X\hat{\beta}) + (Y - X\hat{\beta})'(X\hat{\beta} - X\beta) \\ &\quad + (X\hat{\beta} - X\beta)'(Y - X\hat{\beta}) + (X\hat{\beta} - X\beta)'(X\hat{\beta} - X\beta). \end{aligned}$$

We will show later that the middle two terms are zero. Eliminating them gives

$$(Y - X\beta)'(Y - X\beta) = (Y - X\hat{\beta})'(Y - X\hat{\beta}) + (X\hat{\beta} - X\beta)'(X\hat{\beta} - X\beta), \quad (5)$$

which is easy to minimize. The first term on the right hand side does not depend on  $\beta$ , so the  $\beta$  that minimizes  $(Y - X\beta)'(Y - X\beta)$  is the  $\beta$  that minimizes  $(X\hat{\beta} - X\beta)'(X\hat{\beta} - X\beta)$ . This second term is the sum of squares of the elements in the vector  $X\hat{\beta} - X\beta$ , so it is non-negative and can be minimized by making it zero, which is accomplished by choosing  $\beta = \hat{\beta}$ .

Now let's go back and establish that both the middle terms are 0. They behave similarly, so consider  $(X\hat{\beta} - X\beta)'(Y - X\hat{\beta})$ . From the definition of  $\hat{\beta}$ ,

$$\begin{aligned} (X\hat{\beta} - X\beta)'(Y - X\hat{\beta}) &= [X(\hat{\beta} - \beta)]'(Y - X\hat{\beta}) \\ &= (\hat{\beta} - \beta)'X'(Y - X(X'X)^{-1}X'Y) \\ &= (\hat{\beta} - \beta)'X'(I_n - X(X'X)^{-1}X')Y \end{aligned}$$

but

$$X'(I_n - X(X'X)^{-1}X') = X' - (X'X)(X'X)^{-1}X' = X' - X' = 0,$$

so

$$(\hat{\beta} - \beta)'(Y - X\hat{\beta}) = 0.$$

Similarly  $(Y - X\hat{\beta})'(X\hat{\beta} - X\beta) = 0$ .

EXERCISE 9.2. Use (2) to show that  $E(MSE | \sigma^2) = \sigma^2$ . What is  $\text{Var}(MSE | \sigma^2)$ ?

9.2.2 *Posterior Analysis*

Substituting (5) into the likelihood (9.1.3) gives

$$\begin{aligned} L(\beta, \tau) &\propto \tau^{n/2} \exp \left[ -\frac{\tau}{2} \left\{ (X\hat{\beta} - X\beta)'(X\hat{\beta} - X\beta) + (Y - X\hat{\beta})'(Y - X\hat{\beta}) \right\} \right] \\ &= \tau^{r/2} \exp \left\{ -\frac{\tau}{2} (\hat{\beta} - \beta)'(X'X)(\hat{\beta} - \beta) \right\} \\ &\quad \times \tau^{(n-r)/2} \exp \left\{ -\frac{\tau}{2} (Y - X\hat{\beta})'(Y - X\hat{\beta}) \right\}. \end{aligned}$$

Using the SIR prior  $p(\beta, \tau) = 1/\tau$ , the posterior density can be written

$$\begin{aligned} p(\beta, \tau | Y) &\propto \tau^{r/2} \exp \left\{ -\frac{1}{2} (\hat{\beta} - \beta)'(\tau X'X)(\hat{\beta} - \beta) \right\} \\ &\quad \times \tau^{\frac{n-r}{2}-1} \exp \left\{ -\frac{\tau}{2} (Y - X\hat{\beta})'(Y - X\hat{\beta}) \right\}. \end{aligned}$$

The first term is proportional to a multivariate normal density for  $\beta$  conditional on  $\tau$  and the second term is proportional to a Gamma density for  $\tau$ . In other words,

$$p(\beta, \tau | Y) = p(\beta | \tau, Y)p(\tau | Y)$$

with

$$\begin{aligned} \beta | \tau, Y &\sim N_r \left( \hat{\beta}, (\tau X'X)^{-1} \right) = N_r \left( \hat{\beta}, \sigma^2 (X'X)^{-1} \right); \\ \tau | Y &\sim \text{Gamma} \left( \frac{n-r}{2}, \frac{(Y - X\hat{\beta})'(Y - X\hat{\beta})}{2} \right). \end{aligned}$$

By standardization it follows that

$$\frac{c'\beta - c'\hat{\beta}}{\sqrt{\sigma^2 c'(X'X)^{-1} c}} \Big| \tau, Y \sim N(0, 1)$$

and from Exercise 5.19

$$\frac{(Y - X\hat{\beta})' (Y - X\hat{\beta})}{\sigma^2} \Big| Y = (SSE) \tau \Big| Y \sim \text{Gamma} \left( \frac{n-r}{2}, \frac{1}{2} \right) = \chi_{n-r}^2.$$

Arguments similar to those in Section 5.2 yield

$$\frac{c'\beta - c'\hat{\beta}}{\sqrt{MSE c'(X'X)^{-1} c}} \Big| Y \sim t(n-r)$$

and when predicting a new observation  $y_f$  with covariate vector  $x_f$ ,

$$\frac{y_f - x'_f \hat{\beta}}{\sqrt{MSE [1 + x'_f (X'X)^{-1} x_f]}} \Big| Y \sim t(n-r).$$

Comparing these results to (1), (2), (3), and (4), this prior gives posterior means and posterior probability intervals for linear functions of the regression coefficients and for new observations that agree numerically with the least squares estimates and frequentist confidence intervals, respectively (also see Exercise 9.3). The posterior mean of the precision is  $1/MSE$  and probability intervals

for the variance or precision are numerically equivalent to the frequentist intervals. As far as standard point or interval estimation is concerned, Bayesians have little problem with frequentist normal theory regression results. However, Bayesian testing results are typically quite different from Neyman-Pearson testing results.

Inference for nonlinear functions of regression coefficients, e.g.,  $c'_1\beta/c'_2\beta$ , is more difficult for frequentists and is typically based on large sample approximations. Bayesian estimation is straightforward if it is based on a Monte Carlo sample from the posterior distribution. Simulation from the posterior can be accomplished using the method of composition (cf. Section 3.3) by generating a value  $\tau^k$  from its marginal Gamma posterior and then sampling  $\beta^k$  from the conditional posterior  $N_r(\hat{\beta}, (\tau^k X'X)^{-1})$ . Repeat this two-step sampling process  $m$  times to generate a sample  $\{(\beta^k, \tau^k) : k = 1, \dots, m\}$  from  $p(\beta, \tau | Y)$ .

**EXERCISE 9.3.** *Marginal Posterior of  $\beta$ .* For  $\mu_{r \times 1}$  and positive definite  $\Sigma_{r \times r}$ , define the  $r$ -dimensional (multivariate)  $t_r(n, \mu, \Sigma)$  distribution as having density

$$p_{r,n}(t|\mu, \Sigma) \propto \left[ 1 + \frac{1}{n} (t - \mu)' \Sigma^{-1} (t - \mu) \right]^{-(n+r)/2} [\det(\Sigma)]^{-1/2}.$$

Note that for  $r = 1$ ,  $t(n, \mu, \sigma) = t_1(n, \mu, \sigma^2)$ , cf. Table 2.1. Using arguments similar to those in Subsection 5.2.1, show that

$$\beta | Y \sim t_r \left( n - r, \hat{\beta}, \text{MSE}(X'X)^{-1} \right).$$

**EXERCISE 9.4.** *Marginal Posterior of  $c'\beta$ .* Show that

$$c'\beta | Y \sim t \left( n - r, c'\hat{\beta}, \sqrt{\text{MSE } c'(X'X)^{-1} c} \right).$$

Hint: First obtain the distributional result for  $c'\beta | \tau, Y$  and then obtain the kernel of the marginal density for  $c'\beta | Y$  by integrating the density for  $(c'\beta, \tau) | Y$  with respect to  $\tau$ , cf. Subsection 5.2.1.

**EXERCISE 9.5.** *Prediction of  $y_f$ .* (a) Derive the predictive density for a future response  $y_f$  from an experimental unit that has a known covariate combination  $x_f$ , i.e., show that

$$y_f | x_f, Y \sim t \left( n - r, x_f' \hat{\beta}, \sqrt{\text{MSE } (1 + x_f'(X'X)^{-1} x_f)} \right),$$

cf. Exercise 9.4 and Subsection 5.2.1. (b) Explain how you would use the method of composition to obtain a numerical approximation to a 95% PI for  $y_f$ .

### 9.2.3 A Proper Reference Prior

A commonly used reference prior that is not improper is an approximation to the SIR prior, namely

$$\beta_j \stackrel{iid}{\sim} N(0, b) \quad \perp\!\!\!\perp \quad \tau \sim \text{Gamma}(c, c),$$

where  $b$  is large and  $c$  is small. A standard choice for  $b$  has been  $10^6$  with  $c = 10^{-3}$ . This prior approximates the SIR since the kernel of a  $\text{Gamma}(c, c)$  is  $\tau^{c-1} e^{-\tau c}$ , so for small  $c$  we have  $p(\tau) \propto 1/\tau$ . Similarly,  $p(\beta)$  is approximately uniform when  $b$  is large, cf. Exercise 9.6. The actual values of  $b$  and  $c$  in the normal and gamma distributions may be varied depending on the scale of the measurements being considered. This prior can be used in the primary data analysis or in a sensitivity analysis when an informative prior is available. The posterior analysis must be obtained by simulation. In particular, the MCMC techniques discussed in Section 9.4 apply. The results provided should be very similar to those from the SIR prior.

**EXERCISE 9.6.** Let  $p(u)$  denote the density of  $\beta_j \sim N(0, b)$ . For any two points  $u_1$  and  $u_2$  show that  $p(u_1)/p(u_2)$  converges to 1 as  $b \rightarrow \infty$ .

### 9.3 Conjugate Priors

As in Section 5.2, we do not actually recommend conjugate priors for linear regression because they are hard to elicit (see the end of Subsection 9.4.1). Nonetheless, the mathematical results are pretty, and they have historical and practical significance. The practical significance arises in Bayesian nonparametric analysis involving Dirichlet process mixture models, which are discussed in Chapter 15. Actual computation of the posterior distributions is left as Exercise 9.12 in Subsection 9.4.5. The ideas are similar to those used in Subsection 9.2.2 for the SIR prior but the algebra is much more complex. The additional algebraic tools needed are presented and used in Subsection 9.4.5 to find full conditional distributions for Gibbs sampling. Their application to conjugate priors is messy but straightforward. The bank salary data are reanalyzed at the end of this section.

With sampling distribution

$$Y|\beta, \tau \sim N_n(X\beta, (1/\tau)I_n)$$

the conjugate prior is

$$\beta|\tau \sim N_r(\beta_0, \tau^{-1}C_0), \quad \tau \sim \text{Gamma}(a, b).$$

For reasons discussed later, we actually prefer to specify a prior

$$\tilde{X}\beta|\tau \sim N_r(\tilde{Y}, (1/\tau)D(\tilde{w})),$$

with known  $r \times r$  matrix  $\tilde{X}$ ,  $r$  vector  $\tilde{Y}$ , and diagonal matrix  $D(\tilde{w})$ . We then induce the prior on  $\beta$ . Multiplying  $\tilde{X}\beta$  on the left by  $\tilde{X}^{-1}$ , as mentioned at the end of Section B.2 we get

$$\beta|\tau \sim N_r(\tilde{X}^{-1}\tilde{Y}, (1/\tau)\tilde{X}^{-1}D(\tilde{w})\tilde{X}^{-1}).$$

The two forms for the prior are actually equivalent with

$$\beta_0 = \tilde{X}^{-1}\tilde{Y}, \quad C_0 = \{\tilde{X}'[D(\tilde{w})]^{-1}\tilde{X}\}^{-1}$$

and a somewhat more complicated and non-unique procedure for determining  $\tilde{Y}$ ,  $\tilde{X}$ , and  $\tilde{w}$  from  $\beta_0$  and  $C_0$ .

With the second form of prior for  $\beta$ , in Subsection 9.4.5 we establish that

$$\beta|\tau, Y \sim N_r\left(\hat{\beta}, \tau\left(X'X + \tilde{X}'D^{-1}(\tilde{w})\tilde{X}\right)^{-1}\right), \quad (1)$$

where

$$\hat{\beta} = \left(X'X + \tilde{X}'D^{-1}(\tilde{w})\tilde{X}\right)^{-1} [X'Y + \tilde{X}'D^{-1}(\tilde{w})\tilde{Y}].$$

One can also show that, marginally,

$$\tau|Y \sim \text{Gamma}(Bdfe/2, (Bdfe)BMSE/2)$$

where  $Bdfe = n + 2a$  and

$$BMSE = \frac{(Y - X\hat{\beta})'(Y - X\hat{\beta}) + (\tilde{Y} - \tilde{X}\hat{\beta})'D^{-1}(\tilde{w})(\tilde{Y} - \tilde{X}\hat{\beta}) + 2b}{Bdfe}.$$

One can regard  $2a$  as the number of “prior” observations for  $\tau$  because of its role relative to  $n$  in  $Bdfe$ .

Given these results, arguments similar to those discussed in Section 5.2 and Exercise 9.4 give the marginal posterior

$$\frac{c'\beta - c'\hat{\beta}}{\sqrt{BMSE c'[X'X + \tilde{X}'[D(\tilde{w})]^{-1}\tilde{X}]^{-1}c}} \Big| Y \sim t(Bdfe).$$

For predicting a new observation  $y_f$  with covariate vector  $x_f$ , we get the predictive distribution

$$\frac{y_f - x'_f \hat{\beta}}{\sqrt{BMSE \left( 1 + x'_f [X'X + \tilde{X}'[D(\tilde{w})]^{-1}\tilde{X} ]^{-1}x_f \right)}} \Big| Y \sim t(Bdfe).$$

Much of the analysis can be obtained by making minor changes to the output from a weighted least squares regression program (cf. Bedrick, Christensen, and Johnson, 1996). Using partitioned matrices as discussed in Section A.10 and fitting

$$\begin{bmatrix} Y \\ \tilde{Y} \end{bmatrix} = \begin{bmatrix} X \\ \tilde{X} \end{bmatrix} \beta + \begin{bmatrix} e \\ \tilde{e} \end{bmatrix}, \quad \begin{bmatrix} e \\ \tilde{e} \end{bmatrix} \sim N_{n+r} \left( \begin{bmatrix} 0_{n \times 1} \\ 0_{r \times 1} \end{bmatrix}, \frac{1}{\tau} \begin{bmatrix} I_n & 0 \\ 0 & D(\tilde{w}) \end{bmatrix} \right), \quad (2)$$

one need do little more than modify the reported  $dfe$  and  $MSE$  to agree with the  $Bdfe$  and  $BMSE$ , but that also involves changes to all standard errors.

**EXAMPLE 9.3.1. Bank Salary Data Continued.** We specified five covariate vectors  $\tilde{x}_i' = (1, \tilde{x}_{i2}, \tilde{x}_{i3}, \tilde{x}_{i4}, \tilde{x}_{i5})$  to generate a conjugate prior. Combined, they are

$$\tilde{X} = \begin{bmatrix} \tilde{x}_1' \\ \tilde{x}_2' \\ \tilde{x}_3' \\ \tilde{x}_4' \\ \tilde{x}_5' \end{bmatrix} = \begin{bmatrix} 1 & 0 & 12 & 0 & 0 \\ 1 & 1 & 12 & 0 & 0 \\ 1 & 0 & 15 & 0 & 0 \\ 1 & 0 & 15 & 100 & 0 \\ 1 & 0 & 12 & 0 & 24 \end{bmatrix}.$$

The vector  $\tilde{x}_1' = (1, 0, 12, 0, 0)$  corresponds to a female with 12 years of education, no previous experience, and starting work on January 1, 1969. From this point,  $\tilde{x}_2$  changes the sex,  $\tilde{x}_3$  changes the education,  $\tilde{x}_4$  changes the experience relative to  $\tilde{x}_3$ , and  $\tilde{x}_5$  changes the time hired from  $\tilde{x}_1$ . Thinking about the mean salary for each set of covariates, we chose best guesses for the means of  $\tilde{Y}' = (4000, 4500, 5000, 5500, 5000)$ , which reflects a prior belief that starting salaries are higher for equally qualified men than women (this IS 1977 data remember); a belief that salary is increasing as a function of education, experience, and time; and that three years of education at the beginning of the study has equivalent worth to having been employed at a two-year later date. The weights  $\tilde{w}_i$  are all chosen to be 0.4, so that the prior carries the same weight as  $2 = 5(0.4)$  sampled observations. In other words,

$$\begin{bmatrix} 1 & 0 & 12 & 0 & 0 \\ 1 & 1 & 12 & 0 & 0 \\ 1 & 0 & 15 & 0 & 0 \\ 1 & 0 & 15 & 100 & 0 \\ 1 & 0 & 12 & 0 & 24 \end{bmatrix} \beta \sim N \left( \begin{bmatrix} 4000 \\ 4500 \\ 5000 \\ 5500 \\ 5000 \end{bmatrix}, \frac{1}{\tau} \begin{bmatrix} 0.4 & 0 & 0 & 0 & 0 \\ 0 & 0.4 & 0 & 0 & 0 \\ 0 & 0 & 0.4 & 0 & 0 \\ 0 & 0 & 0 & 0.4 & 0 \\ 0 & 0 & 0 & 0 & 0.4 \end{bmatrix} \right).$$

If we combine this prior for  $\beta$  given  $\tau$  with an improper prior

$$p(\tau) = 1/\tau,$$

posterior results are as if  $a = b = 0$  in the Gamma prior. The posterior results in Table 9.2 were obtained using a weighted least squares regression analysis on the 93 observations in the original data taken together with the 5 weighted prior observations as in model (2) using weights from the diagonal elements of the covariance matrix in (2).

For everything except  $\tau$ , the tabulated means and scales are exactly as in weighted least squares output. As in Example 9.2.1, we report posterior scale parameters for the  $t(93)$  distributions rather than the posterior standard deviations. The posterior 95% probability intervals are obtained from  $\text{Mean} \pm t(0.975, Bdfe) \text{ Scale}$  where  $t(0.975, 93) = 1.9858$ . Again  $x'_f = (1, 1, 16, 54.5, 33.0)$  corresponds to the highest paid individual in the data, a man making \$8,100.

Table 9.2: Regression on bank salary data with an informative prior for  $\beta$ .

Parameter	Mean	Scale	2.5%	97.5%
Constant	3470.4	320.1	2834.7	4106.0
Sex	708.5	114.4	481.4	935.7
Educ	93.35	24.08	45.54	141.17
Exp	1.3494	0.5725	0.21	2.49
Time	23.836	5.004	13.90	33.77
$x'_f \beta$	6532.7	142.9	6248.9	6816.5
$y_f$	6532.7	519.3	5501.4	7564.0
$\tau$	$4.01 \times 10^{-6}$	$0.588 \times 10^{-6}$	$2.94 \times 10^{-6}$	$5.24 \times 10^{-6}$

Inferences are similar to those in Example 9.2.1. There are slight changes in estimated regression coefficients. Posterior scale factors are uniformly slightly smaller here. The implication is that the data and the prior information are not obviously inconsistent with one another and that the prior information actually reduces posterior uncertainty, although only slightly.

Inference about  $\sigma$  and  $\tau$  is similar to the SIR analysis. The weighted least squares analysis gives  $Bdfe = 93$  and  $BSSE = 23,181,306$ , so  $BMSE = 249,261$  and  $\hat{\tau} = E(\tau|Y) = 1/(249,261)$ . Percentiles of  $\tau$  are obtained by dividing the percentiles of a  $\chi^2_{93}$  by  $BSSE$ , where  $\chi^2_{93}(0.025) = 68.2112$  and  $\chi^2_{93}(0.975) = 121.571$ . The 95% PI for  $\sigma$  is  $1000(5.24^{-0.5}, 2.94^{-0.5}) = (436.9, 583.2)$ , which is practically the same as the SIR result.

The squared correlation between the 93 actual observations and their predicted values is  $R^2 = 0.51$ . Such a definition of  $R^2$  is appropriate for most prediction problems.

Finally, we take a proper conjugate prior on the precision,

$$\tau \sim \text{Gamma}(1, 250000).$$

This is based on making the prior worth 2 prior observations on  $\tau$  with a mean of 1/250000, and which roughly corresponds to thinking that the standard deviation  $\sigma$  is 500. A proper gamma prior on  $\tau$  does not affect the posterior mean of linear functions  $c'\beta$  or new observations  $y_f$ , but a proper gamma prior on  $\tau$  might have an appreciable effect on posterior standard deviations and percentiles. In our example, it does not. In our example, the new  $Bdfe$  is  $Bdfe = 95$  with  $BSSE = 23,181,306 + 2(250,000) = 23,681,306$  and  $BMSE = 249,277$ . The new scale factors are the factors reported by weighted least squares multiplied by  $\sqrt{249,277/249,261} = 1.0001$ , i.e., the scale factor for Sex with this prior on  $\tau$  will be  $1.0001(114.4) = 114.4$ .

We have been somewhat cavalier about the choice of prior for  $\tau$ , although the rough guess of 500 for  $\sigma$  turned out to be squarely in the middle of the posterior PI for it. In the next section, we incorporate a more sophisticated elicitation of prior variability. We explain why we prefer not to elicit conjugate priors at the end of Subsection 9.4.1 after we have explained our preferred method.

## 9.4 Independence Priors

This section presents the priors that we use most often in practice. We assume prior independence of  $\beta$  and  $\tau$ , i.e.,  $p(\beta, \tau) = p(\beta)p(\tau)$ . We also discuss model fitting in WinBUGS and R, and inferences for parameters, functions of parameters (e.g., relative means), and future values. We begin with a description of data that are discussed throughout the remainder of the chapter.

**EXAMPLE 9.4.1. FEV Data.** In Chapter 7 we considered data from Rosner (2006) on pulmonary function (lung capacity) in adolescents. The response  $y$  is forced expiratory volume (FEV), which measures the volume of air in liters expelled in 1 second of a forceful breath. Lung function is expected to increase during adolescence, but smoking may slow its progression. The association

between smoking and lung capacity in adolescents is investigated using data from 345 adolescents between the ages of 10 and 19. The predictor variables include a constant, age in years ( $x_2$ ), a 0/1 indicator variable for smoking status ( $x_3$ ), and the interaction term  $x_4 = x_2x_3$ . A predictor vector is  $x' = (1, x_2, x_3, x_4)$ . The data were presented in Figure 7.2.

With sampling model

$$Y | \beta, \tau \sim N_n(X\beta, \tau^{-1}I_n),$$

our standard informative prior has independent components

$$\beta \sim N_r(\beta_0, C_0) \quad \perp \!\!\! \perp \quad \tau \sim \text{Gamma}(a, b). \quad (1)$$

Our task is to determine values  $a$ ,  $b$ ,  $\beta_0$ , and  $C_0$  that accurately reflect available prior information. For  $\beta$ , we actually prefer to specify an equivalent prior

$$\tilde{X}\beta \sim N_r(\tilde{Y}, D(\tilde{w})), \quad (2)$$

with known  $r$ -vector  $\tilde{Y}$ , nonsingular  $r \times r$  matrix  $\tilde{X}$ , and diagonal matrix  $D(\tilde{w})$ . We then induce the prior on  $\beta$ . As with the conjugate prior, multiply  $\tilde{X}\beta$  on the left by  $\tilde{X}^{-1}$ . This now gives

$$\beta \sim N_r(\tilde{X}^{-1}\tilde{Y}, \tilde{X}^{-1}D(\tilde{w})\tilde{X}^{-1}). \quad (3)$$

The BCJ method of constructing an informative prior for  $\beta$  as embodied in (2) and (3) is analogous to the method used in Section 8.4 for logistic regression and similar to that used in the previous section. Our goal is not to find the perfect prior to characterize someone's beliefs but to find a sensible prior that incorporates some basic scientific or experiential knowledge. We frame our discussion of prior elicitation around the FEV study.

#### 9.4.1 Prior on $\beta$

Individual regression coefficients  $\beta_j$  are difficult to think about, since they are only indirectly related to anything observable. Instead of specifying a prior directly on  $\beta$ , we specify a prior on values that are more closely related to observables and induce the prior for  $\beta$ .

It is relatively easy to elicit a prior for the mean value of something that is potentially observable. We elicit a prior for the mean of the observations corresponding to a particular set of predictor variables. Technically, we elicit priors on conditional means

$$\tilde{m}_i \equiv E[y | \tilde{x}_i] = \tilde{x}'_i \beta$$

for  $r$  subpopulations defined by different predictor vectors  $\tilde{x}_i$ ,  $i = 1, \dots, r$ . The prior has the form (2) and the induced prior on  $\beta$  has the form (3).

With four regression parameters in the FEV model, we pick four covariate combinations  $\tilde{x}_i$ ,  $i = 1, 2, 3, 4$ , corresponding to circumstances that the expert can assess independently. In particular, we use prior experience to specify information about the subpopulation of 11-year-old nonsmokers,  $\tilde{x}'_1 = (1, 11, 0, 0)$ ; 13-year-old smokers,  $\tilde{x}'_2 = (1, 13, 1, 13)$ ; 16-year-old nonsmokers,  $\tilde{x}'_3 = (1, 16, 0, 0)$ , and 18-year-old smokers,  $\tilde{x}'_4 = (1, 18, 1, 18)$ . In matrix form we have

$$\tilde{X} = \begin{bmatrix} 1 & 11 & 0 & 0 \\ 1 & 13 & 1 & 13 \\ 1 & 16 & 0 & 0 \\ 1 & 18 & 1 & 18 \end{bmatrix} = \begin{bmatrix} \tilde{x}'_1 \\ \tilde{x}'_2 \\ \tilde{x}'_3 \\ \tilde{x}'_4 \end{bmatrix}.$$

FEV values display variability within subpopulations. We are interested in the mean FEV value  $\tilde{m}_i$  for the four circumstances  $\tilde{x}_1, \tilde{x}_2, \tilde{x}_3$ , and  $\tilde{x}_4$ , and ultimately in placing normal priors on the mean FEVs for the four types of adolescents defined by these covariate combinations. The elicitation

process mimics that for the one- and two-sample normal cases discussed in Section 5.2, only here we are eliciting information about four groups.

As an example, suppose that our medical collaborator expects the average FEV among all 18-year-old smokers in the sampled population to be 3.3, and is 99% sure that the mean FEV is less than 4.0 in this group. We take  $\tilde{y}_4 = 3.3$  to be the mean of  $\tilde{m}_4$ . The prior variance reflects our uncertainty about where the mean lies, and if a normal prior is assumed then the 99th percentile of  $\tilde{m}_4$  satisfies  $4.0 = 3.3 + 2.33\sqrt{\tilde{w}_4}$ , so  $\tilde{w}_4 = (4.0 - 3.3)^2/2.33^2 = 0.09$  and

$$\tilde{m}_4 \sim N(3.3, 0.09).$$

Similar steps are used to construct normal priors for  $\tilde{m}_1, \tilde{m}_2$ , and  $\tilde{m}_3$ ; suppose they yield

$$\tilde{m}_1 \sim N(2.8, 0.04),$$

$$\tilde{m}_2 \sim N(3.0, 0.04),$$

$$\tilde{m}_3 \sim N(4.0, 0.04).$$

It is important to choose the  $\tilde{x}_i$ s so that the  $\tilde{m}_i$ s can be regarded independently. Noting that  $\tilde{m}_i = \tilde{x}'_i \beta$ , we have a prior of the form (2),

$$\tilde{X}\beta \equiv \tilde{m} = \begin{bmatrix} \tilde{m}_1 \\ \tilde{m}_2 \\ \tilde{m}_3 \\ \tilde{m}_4 \end{bmatrix} \sim N \left( \begin{bmatrix} 2.8 \\ 3.0 \\ 4.0 \\ 3.3 \end{bmatrix}, \begin{bmatrix} 0.04 & 0 & 0 & 0 \\ 0 & 0.04 & 0 & 0 \\ 0 & 0 & 0.04 & 0 \\ 0 & 0 & 0 & 0.09 \end{bmatrix} \right).$$

We need  $p(\beta)$ , the prior density on  $\beta$ . We have a prior on  $\tilde{m}$  with, say, density  $q(\tilde{m})$ . Since  $\tilde{m} = \tilde{X}\beta$ , we have  $\beta = \tilde{X}^{-1}\tilde{m}$  and we can infer the distribution on  $\beta$ . One way to do this is to use the change of variable formula for densities, Proposition B.4. An easier way to do it is to use the fact that since  $\tilde{m}$  has a multivariate normal distribution,  $\beta = \tilde{X}^{-1}\tilde{m}$  will also have a multivariate normal distribution with appropriately transformed mean vector and covariance matrix. In particular,

$$\beta = \tilde{X}^{-1} \begin{bmatrix} \tilde{m}_1 \\ \tilde{m}_2 \\ \tilde{m}_3 \\ \tilde{m}_4 \end{bmatrix} \sim N \left( \tilde{X}^{-1} \begin{bmatrix} 2.8 \\ 3.0 \\ 4.0 \\ 3.3 \end{bmatrix}, \tilde{X}^{-1} \begin{bmatrix} 0.04 & 0 & 0 & 0 \\ 0 & 0.04 & 0 & 0 \\ 0 & 0 & 0.04 & 0 \\ 0 & 0 & 0 & 0.09 \end{bmatrix} \tilde{X}^{-1'} \right)$$

where

$$\tilde{X}^{-1} = \begin{bmatrix} 3.2 & 0 & -2.2 & 0 \\ -0.2 & 0 & 0.2 & 0 \\ -3.2 & 3.6 & 2.2 & -2.6 \\ 0.2 & -0.2 & -0.2 & 0.2 \end{bmatrix},$$

so

$$\beta \sim N \left( \begin{bmatrix} 0.16 \\ 0.24 \\ 2.06 \\ -0.18 \end{bmatrix}, \begin{bmatrix} 0.603 & -0.043 & -0.603 & 0.043 \\ -0.043 & 0.003 & 0.043 & -0.003 \\ -0.603 & 0.043 & 1.73 & -0.119 \\ 0.043 & -0.003 & -0.119 & 0.008 \end{bmatrix} \right).$$

**EXERCISE 9.7.** Consider the prior specified by

$$\tilde{X} = \begin{bmatrix} 1 & 11 & 0 & 0 \\ 1 & 13 & 1 & 13 \\ 1 & 16 & 0 & 0 \\ 1 & 18 & 1 & 18 \end{bmatrix}, \quad \begin{bmatrix} \tilde{m}_1 \\ \tilde{m}_2 \\ \tilde{m}_3 \\ \tilde{m}_4 \end{bmatrix} \sim N \left( \begin{bmatrix} 2.3 \\ 3.0 \\ 4.0 \\ 2.8 \end{bmatrix}, \begin{bmatrix} 0.04 & 0 & 0 & 0 \\ 0 & 0.04 & 0 & 0 \\ 0 & 0 & 0.04 & 0 \\ 0 & 0 & 0 & 0.09 \end{bmatrix} \right).$$

Find the prior distribution for  $\beta$  and discuss the differences between this prior and the one just illustrated.

**EXERCISE 9.8.** Consider fitting a simple linear regression model of college grade point average (CGPA) on high school grade point average (HSGPA). Suppose you elicit a prior for  $\tilde{m}_1$ , the average CGPA for students with an HSGPA of 3.0, that has a prior guess of 2.8, and you are 95% sure that the average CGPA for such high school students is less than 3.2. Moreover, suppose for students with HSGPA = 3.8, your best guess for  $\tilde{m}_2$  is 3.5 with a 95% upper limit of 3.8. (a) Derive the induced BCJ prior on  $\beta$  and give 95% prior PIs for  $\beta_1$  and  $\beta_2$ . (b) Simulate the induced prior on  $\beta$  in WinBUGS and compare 95% prior probability intervals with those in (a). (c) Place independent uniform priors on the two mean CGPAs, say  $\tilde{m}_1 \sim U(2.3, 3.5)$  and  $\tilde{m}_2 \sim U(2.8, 3.9)$ . Use WinBUGS to obtain the new induced prior on  $\beta$  and compare with the previous results.

In Section 9.3 on conjugate priors we also used an  $\tilde{X}$  to elicit prior means  $\tilde{Y}$  but there we treated the  $\tilde{w}_i$ s rather cavalierly. We feel that the independent priors approach makes it easier to elicit realistic information about variability. With a conjugate prior, information about variability related to regression coefficients must be conditional on  $\tau$ . For example, when asking the expert for, say, a 99th percentile that reflects uncertainty about the mean FEV for 18-year-old smokers, that value must be conditional on the variance  $\sigma^2$ . If we tell the expert that the variance of the data is 2.0, and the expert says the 99% upper bound for  $\tilde{m}_4$  is 4.0, we get  $4.0 = 3.3 + 2.33\sigma\sqrt{\tilde{w}_4} = 3.3 + 2.33\sqrt{2.0\tilde{w}_4}$  so  $\tilde{w}_4 = (4.0 - 3.3)^2/[2.33^2(2.0)] = 0.045$ . But to be consistent with this conjugate model, the expert would have to be willing to specify an upper bound of  $5.7 = 3.3 + 2.33\sqrt{4.0(0.045)}$ , if we told her that the data variance was 4 instead of 2. We find that asking for bounds conditional on the variance makes experts uncomfortable.

#### 9.4.2 Prior on $\tau$

We now construct an informative prior on  $\tau$  or  $\sigma$ . The approach is a slight modification of that presented in Subsection 5.2.3. We continue to illustrate with the FEV example. To choose  $a$  and  $b$  for a  $\text{Gamma}(a, b)$  prior on  $\tau$  we again consider 18-year-old smokers,  $\tilde{x}'_4 = (1, 18, 1, 18)$ . We now elicit the largest *observation* the expert would expect to see from this group (as opposed to the largest value the mean could be), given our best guess for  $\tilde{m}_4$ . This provides a best guess for the 95th percentile of FEVs among 18-year-old smokers. Under normality, the 95th percentile is  $\tilde{m}_4 + 1.645\sigma$ , where 1.645 is the 95th percentile of the  $N(0, 1)$  distribution. But the best guess is conditional on  $\tilde{m}_4 = \tilde{y}_4$ , so with the expert's best guess for the 95th percentile being 5 and with  $\tilde{y}_4 = 3.3$ , our best guess for  $\sigma$ , say  $\sigma_0$ , satisfies  $5 = 3.3 + 1.645\sigma_0$ . It follows that the best guess for  $\sigma$  is  $\sigma_0 = 1.7/1.645$  or  $\tau_0 = (1.645/1.7)^2 = 0.94$  for  $\tau$ . We take  $\tau_0$  to be the mode of the gamma prior for  $\tau$ , that is

$$0.94 = \frac{a-1}{b}$$

or, solving for  $a$ ,

$$a(b) = 0.94b + 1.$$

It might be worthwhile to see whether the expert believes the sampling model we are proposing. Suppose we again consider 18-year-old smokers, but assume now that the best guess for their mean FEV is 3.5 (instead of 3.3). If the expert thinks that the 95th percentile is substantially different from 5.2, they think the precision depends on the mean value, which is contrary to our model. Moreover, if we consider, say, 11-year-old nonsmokers and a conditional mean FEV of 2.8, to be consistent with the regression model, their corresponding 95th percentile should be about 4.5. According to the model, to be consistent with the first information, the 95th percentile should be about 1.7 units above the mean FEV, regardless of the values of the predictor variables  $x_2$ ,  $x_3$ , and  $x_4$ .

To complete our specification of the gamma prior, we elicit an uncertainty about the 95th percentile. In addition to a best guess for the 95th percentile, we give an operative lower bound to it by specifying that we are 95% certain that the 95th percentile is greater than 4. It follows that

$$\begin{aligned} 0.95 &= \Pr(\mu + 1.645\sigma > 4 | \mu = 3.3) \\ &= \Pr(3.3 + 1.645\sigma > 4 | \mu = 3.3) \end{aligned}$$

$$\begin{aligned}
&= \Pr(1.645\sigma > 0.7) \\
&= \Pr(\sigma > 0.7/1.645) \\
&= \Pr(\tau < (1.645/0.7)^2) \\
&= \Pr(\tau < 5.52).
\end{aligned}$$

Take a grid of  $b$  values and find the  $\text{Gamma}(a(b), b)$  distribution that has 95th percentile 5.52. We used the following R commands to find the prior parameters:

```

b <- seq(0,1,0.01)
a <- 0.94*b+1
cbind(a,b,pgamma(5.52,a,b)) # Look for 0.95

```

The output of the program includes

```

[78,] 1.7238 0.77 0.94832227
[79,] 1.7332 0.78 0.95004899
[80,] 1.7426 0.79 0.95171579

```

and we see that

$$\tau \sim \text{Gamma}(1.73, 0.78)$$

approximately satisfies these requirements.

We can similarly induce a Gamma prior on  $\sigma$  by fixing the mode and a percentile as above. See Subsection 5.2.3 for additional details.

**EXERCISE 9.9. *GPA Data.*** For the problem described in Exercise 9.8, suppose that the 90th percentile of CGPA s among students with a 3.0 HSGPA was thought to be 3.5. Moreover, with 95% certainty assume that the 90th percentile is at least 3.2. Derive an appropriate Gamma prior for  $\sigma$  given this information.

#### 9.4.3 Partial Prior Information

In the absence of substantive prior knowledge, the SIR prior will likely suffice. However, if substantive prior information is available, we would be loathe to ignore it. Sometimes the available information is insufficient to specify a complete BCJ prior on  $r$  regression coefficients, especially when  $r$  is moderate or large. It is not a simple matter to place a BCJ prior on even 6 regression coefficients, as was seen in Chapter 8 with the trauma data. Eliciting real prior information can be time consuming. This subsection discusses how to place a partial prior on regression coefficients when we are only able to specify information for  $r_1 < r$  mean values corresponding to  $r_1$  predictor vectors  $\tilde{x}_i$ .

The simplest case is  $r_1 = 1$  and prior information is available for one “typical” individual in the population. For models with an intercept, consider an individual whose dichotomous predictor variables were all equal to 0 with all the continuous covariate values set to their sample mean from the data. If the continuous variates are standardized, then we simply set  $\tilde{x}_1 = (1, 0, \dots, 0)'$ . In this case,  $\tilde{m}_1 = \beta_1$ , i.e., the intercept. We place an informative prior normal distribution on  $\tilde{m}_1$  as in Subsection 9.4.1. We now require some kind of a reference prior to be placed on  $(\beta_2, \dots, \beta_r)$ . We either pick an improper flat prior for these coefficients or a set of independent normal priors with mean 0 and large variances.

If an additional elicitation is made for  $\tilde{m}_2$  corresponding to predictor vector  $\tilde{x}_2 = (1, 1, 0, \dots, 0)$ , then independent normal distributions are placed on the  $\tilde{m}_i$ s, which induces a joint normal distribution on  $\beta_1$  and  $\beta_2$  via

$$\begin{pmatrix} \tilde{m}_1 \\ \tilde{m}_2 \end{pmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}.$$

A reference prior can be placed on the remaining  $\beta_j$ s.

Formally, let  $\tilde{X}$  be an  $r_1 \times r$  matrix containing the row vectors  $\tilde{x}_i'$ , and let  $\tilde{m} = \tilde{X}\beta$  be the  $r_1 \times 1$  vector of mean responses that correspond to  $\tilde{X}$ . We would generally pick  $\tilde{X}$  so that  $\tilde{m}$  only depends on  $r_1$  of the  $\beta_j$ 's, although this is not strictly necessary, cf. Subsection 8.4.5. With our recommended selection of

$$\tilde{X} = (\tilde{X}_1, \tilde{X}_2) = (\tilde{X}_1, 0),$$

$\tilde{X}_1$  is  $r_1 \times r_1$ . The rows of  $\tilde{X}$  must be linearly independent, as was the case with the full BCJ prior, so  $\tilde{X}_1$  is nonsingular. Write  $\tilde{\beta} = (\beta_1, \dots, \beta_{r_1})'$  so

$$\tilde{m} = \tilde{X}\beta = \tilde{X}_1\tilde{\beta},$$

which induces a prior on  $\tilde{\beta}$ . Subsection 8.4.5 applies directly by letting  $F$  be the identity transformation and  $\tilde{m} = \tilde{\theta}$ , thus we are justified in placing independent  $N(0, b)$  priors on the remaining parameters  $(\beta_{r_1+1}, \dots, \beta_r)$ .

**EXAMPLE 9.4.2.** *FEV Data.* There are two additional predictor variables available for the FEV data: height and sex. In the expanded regression problem we can use the BCJ prior of Subsection 9.4.1 as a partial information prior in conjunction with diffuse priors on the two additional regression coefficients. This presumes that the earlier elicitation applies to children in the four subpopulations with, say, the same average height and the same sex. This is somewhat unrealistic in that adding knowledge of a person's sex would probably change ones opinions about 18 year-old smokers mean FEV. In any case, the parametrization of the model requires height to be standardized and that the selected sex correspond to a predictor value of 0.

**EXERCISE 9.10.** *GPA Data.* Now suppose that there is a second predictor variable for the GPA data, say the Scholastic Assessment Test (SAT) score. The SAT is given in high school and used as a predictor of success in college. (a) Using the prior developed in Exercise 9.8(a), construct a partial prior for regressing CGPA on both HSGPA and SAT. (b) Write WinBUGS code to simulate from your partial prior and induce a prior on mean values corresponding to four individuals, two with 3.0 HSGPAs where one student has a standardized SAT score of 0 (average score on unstandardized scale) and the other has a standardized SAT score of 2 (two standard deviations above the mean), and two others for students with HSGPA of 3.8 and SAT scores of 0 and 2, respectively. Comment.

#### 9.4.4 Inference and Displays

As seen in the next subsection, a Gibbs sampler generates values from  $p(\beta, \tau | Y)$  by iteratively simulating the precision from a gamma distribution and simulating the vector of regression coefficients from a multivariate normal distribution. Posterior estimates for any function  $g(\beta, \tau)$  are computed in the usual way using the post burn-in MCMC sample  $\{(\beta^k, \tau^k) : k = 1, \dots, m\}$ . Examples include estimating subpopulation means  $x'\beta$  for known  $x$  using the sample  $\{x'\beta^k\}_{k=1}^m$ , estimating the difference between two subpopulation means  $x'\beta - \check{x}'\beta$  using the sample  $\{x'\beta^k - \check{x}'\beta^k\}_{k=1}^m$ , or estimating the ratio of two subpopulation means  $x'\beta / \check{x}'\beta$  using  $\{x'\beta^k / \check{x}'\beta^k\}_{k=1}^m$ . For each case we can compute posterior means, standard deviations, medians, and percentiles to give probability intervals. Predictive densities and point and interval predictions are readily obtained by sampling future observations  $y_f^k$  with covariate vector  $x_f$  from the normal sampling model given the current  $\beta^k$  and  $\tau^k$ .

Tables can be used to assess the relative importance of predictors. For regressing, say, CGPA on HSGPA, standardized SAT scores, and Sex, similar to Exercises 8, 9, and 10, consider the 8 conditions consisting of all combinations of males and females, HSGPA scores of 3.0 and 3.5, and standardized SAT scores of 0 and 2. In other words, consider the 8 covariate vectors

(1, 3.0, 0, 0)	(1, 3.0, 2, 0)
(1, 3.0, 0, 1)	(1, 3.0, 2, 1)
(1, 3.5, 0, 0)	(1, 3.5, 2, 0)
(1, 3.5, 0, 1)	(1, 3.5, 2, 1).

For each, find the posterior median and 95% PI for  $x'\beta$ . The results might be summarized as:

		SAT = 0		SAT = 2	
		Med	95% PI	Med	95% PI
HSGPA = 3.0	Male	2.81	(2.61, 3.01)	3.00	(2.80, 3.20)
	Female	2.80	(2.60, 3.00)	3.01	(2.81, 3.21)
HSGPA = 3.5	Male	3.29	(3.09, 3.49)	3.48	(3.28, 3.68)
	Female	3.32	(3.12, 3.52)	3.51	(3.31, 3.71)

We can see that the effect of increasing SAT score from average to two standard deviations above the mean results in an estimated increase in average CGPA of about 0.2, regardless of the gender or HSGPA. Increasing HSGPA from 3.0 to 3.5 results in an estimated increase in mean CGPA of about a half grade point, regardless of gender or SAT score, and there is no apparent effect of gender on CGPA. To establish the statistical import of the interpretations drawn from the table, we need to look at corresponding posterior probabilities. To obtain these interpretations, the PI for the HSGPA regression coefficient should be tightly packed near 1 and the PI for the SAT coefficient should be tightly packed about 0.1. We also use posterior PIs on regression coefficients to see if the corresponding variables are important. They can be deemed important if the PIs are removed from 0. Having the PI for the Sex regression coefficient narrow and including zero supports our conclusion that Sex has little association with CGPA.

As an alternative to the table, we could plot the estimated mean CGPA as a function of SAT score for both males and females, with HSGPA fixed at 3.0 on one plot and fixed at 3.5 on a second plot. Similarly, we could plot the estimated mean CGPA as a function of HSGPA for SAT scores of 0 and 2 on a single plot for males, and a similar plot for females. Conclusions drawn from these measures of central tendency should be supported by appropriate probability intervals. Chapter 8 gives similar plots for the trauma data analysis. Appendix C provides R examples showing how to produce side-by-side plots and graphs that contain multiple curves.

#### 9.4.5 Gibbs Sampling\*

For normal-gamma independence priors, the posterior distribution  $p(\beta, \tau | Y)$  is analytically intractable, but the full conditionals needed for Gibbs sampling are standard distributions. Our discussion makes extensive use of partitioned matrices as presented in Section A.10.

For linear regression based on the likelihood (9.1.3) and the independence prior specified by (1), (2), and (3), we have

$$\begin{aligned} p(\beta, \tau | Y) &= L(\beta, \tau)p(\tau)p(\beta) \\ &\propto \tau^{n/2} \exp\left\{-\frac{\tau}{2}(Y - X\beta)'(Y - X\beta)\right\} \\ &\quad \times \tau^{a-1} \exp(-b\tau) \exp\left\{-\frac{1}{2}(\tilde{X}\beta - \tilde{Y})' D(\tilde{w})^{-1} (\tilde{X}\beta - \tilde{Y})\right\}. \end{aligned}$$

Dropping multiplicative terms that do not involve  $\tau$ , the full conditional for  $\tau$  has density

$$p(\tau | \beta, Y) \propto \tau^{\frac{n+2a}{2}-1} \exp\left[-\frac{\tau}{2}\{2b + (Y - X\beta)'(Y - X\beta)\}\right],$$

so

$$\tau | \beta, Y \sim \text{Gamma}\left(\frac{n+2a}{2}, \frac{1}{2}\{2b + (Y - X\beta)'(Y - X\beta)\}\right).$$

Because Gibbs sampling involves sampling repeatedly from this distribution for successive  $\beta$  iterates, it may be more computationally efficient to write

$$(Y - X\beta)'(Y - X\beta) = (Y - X\hat{\beta})'(Y - X\hat{\beta}) + (\hat{\beta} - \beta)'(X'X)(\hat{\beta} - \beta)$$

with  $\hat{\beta}$  the least squares estimate. The first term on the right-hand side is a constant in  $\beta$  so the repeated computations involve only the second term, which consists of  $r$  dimensional matrix products rather than  $n$  dimensional products.

To obtain the full conditional for  $\beta$ , rewrite the joint density as

$$p(\beta, \tau, Y) \propto \tau^{n/2} \tau^{a-1} \exp(-b\tau) \times \\ \exp \left\{ -\frac{1}{2} (X\beta - Y)' (\tau^{-1} I_n)^{-1} (X\beta - Y) - \frac{1}{2} (\tilde{X}\beta - \tilde{Y})' D(\tilde{w})^{-1} (\tilde{X}\beta - \tilde{Y}) \right\}.$$

Dropping multiplicative terms that do not involve  $\beta$  gives

$$p(\beta | \tau, Y) \propto \exp \left\{ -\frac{1}{2} \begin{bmatrix} X\beta - Y \\ \tilde{X}\beta - \tilde{Y} \end{bmatrix}' \begin{bmatrix} \tau^{-1} I_n & 0 \\ 0 & D(\tilde{w}) \end{bmatrix}^{-1} \begin{bmatrix} X\beta - Y \\ \tilde{X}\beta - \tilde{Y} \end{bmatrix} \right\}$$

or

$$p(\beta | \tau, Y) \propto \exp \left\{ -\frac{1}{2} \left( \begin{bmatrix} X \\ \tilde{X} \end{bmatrix} \beta - \begin{bmatrix} Y \\ \tilde{Y} \end{bmatrix} \right)' \begin{bmatrix} \tau^{-1} I_n & 0 \\ 0 & D(\tilde{w}) \end{bmatrix}^{-1} \left( \begin{bmatrix} X \\ \tilde{X} \end{bmatrix} \beta - \begin{bmatrix} Y \\ \tilde{Y} \end{bmatrix} \right) \right\}. \quad (4)$$

Define

$$\tilde{\beta} \equiv [\tau X'X + \tilde{X}'D(\tilde{w})^{-1}\tilde{X}]^{-1} (\tau X'Y + \tilde{X}'D(\tilde{w})^{-1}\tilde{Y}).$$

Using Exercise 9.11(a), some matrix algebra, and dropping the term that does not involve  $\beta$ , the posterior density becomes

$$p(\beta | \tau, Y) \propto \exp \left\{ -\frac{1}{2} (\beta - \tilde{\beta})' [\tau X'X + \tilde{X}'D(\tilde{w})^{-1}\tilde{X}] (\beta - \tilde{\beta}) \right\}, \quad (5)$$

which implies

$$\beta | \tau, Y \sim N_r \left( \tilde{\beta}, [\tau X'X + \tilde{X}'D(\tilde{w})^{-1}\tilde{X}]^{-1} \right).$$

Note the similarity of the full conditional for  $\beta$  to (9.3.1). Exercise 9.12 includes deriving (9.3.1) using the methods developed here. This full conditional could also have been obtained by using the complete the square formula, see Exercise 9.11(c).

To obtain the full conditionals for a prior of form (1) that is not of form (3), e.g., a proper reference prior, substitute  $\tilde{X}'D(\tilde{w})^{-1}\tilde{X} = C_0^{-1}$  and  $\tilde{X}'D(\tilde{w})^{-1}\tilde{Y} = C_0^{-1}\beta_0$ .

**EXERCISE 9.11.** (a) Show that

$$(Y - X\beta)'V^{-1}(Y - X\beta) = (Y - X\hat{\beta})'V^{-1}(Y - X\hat{\beta}) + (\hat{\beta} - \beta)'[X'V^{-1}X](\hat{\beta} - \beta)$$

where  $\hat{\beta} = [X'V^{-1}X]^{-1}X'V^{-1}Y$ , cf. the similar result in Subsection 9.2.1. (b) Applying part (a) to the partitioned matrices in (4), show (5). (c) Alternatively, apply the formula from (a) to establish the “complete the squares formula”

$$\begin{aligned} & (\beta - \beta_1)'C_1(\beta - \beta_1) + (\beta - \beta_2)'C_2(\beta - \beta_2) \\ &= \left( \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} - \begin{bmatrix} I_r \\ I_r \end{bmatrix} \beta \right)' \left( \begin{bmatrix} C_1 & 0 \\ 0 & C_2 \end{bmatrix} \right) \left( \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} - \begin{bmatrix} I_r \\ I_r \end{bmatrix} \beta \right) \\ &= (\beta - \hat{\beta})'[C_1 + C_2](\beta - \hat{\beta}) + (\beta_1 - \hat{\beta})'C_1(\beta_1 - \hat{\beta}) + (\beta_2 - \hat{\beta})'C_2(\beta_2 - \hat{\beta}) \end{aligned}$$

with

$$\hat{\beta} = [C_1 + C_2]^{-1}(C_1\beta_1 + C_2\beta_2)$$

and use this to establish (5).

**EXERCISE 9.12.** Modify the ideas of this subsection and Subsection 9.2.2 to derive the posterior distribution for the conjugate prior of Section 9.3.

#### 9.4.6 WinBUGS and R Code

In this subsection we give WinBUGS code for analyzing Example 9.4.1 on FEV using the independence priors developed in Subsections 9.4.1 and 9.4.2. We also add a new twist to the computing that involves driving WinBUGS from within R, cf. Section C.6.

In WinBUGS, the prior on  $\tau$  is a familiar gamma distribution, while the prior on  $\beta$  is specified as in (1) as a multivariate normal distribution with mean vector  $\beta_0$ , and a precision matrix. The precision matrix  $C_0^{-1}$  is the inverse of the covariance matrix  $C_0$ . We have had little luck computing the precision matrix in WinBUGS so we assume that it has been computed elsewhere and read into WinBUGS.

Alternatively, we can run an R program to compute the inverse and call the WinBUGS analysis from the same R program. This avoids the pointing and clicking of interactive WinBUGS. It also avoids the annoyance of finding the precision matrix and hauling it over to WinBUGS, formatted appropriately. Moreover, when running similar WinBUGS programs, for example when hyperparameters of priors are modified in a sensitivity analysis or when different predictor variables are used in a sequence of regression analyses, it is often simpler to modify the R program. This is especially useful when  $C_0$  changes, requiring us to compute additional inverses to obtain the precision matrices.

The WinBUGS code defines a number of inferential objects of interest. These include the estimated mean FEV for smokers and nonsmokers aged 10 to 19, some relative means and differences in means, and predictions of FEV for two types of individuals. The data (FEV, Age, and Smoke) and constants ( $n$ ,  $r$ ,  $a$ ,  $b$ ,  $\beta_0$ , and  $C_0^{-1}$ ) must be supplied to WinBUGS. All will change when using this code as a pattern for other analyses. Later we provide R code that both supplies this information and executes the WinBUGS program.

```
model{
  for(i in 1:n){# The likelihood is specified in this 'for loop'
    FEV[i] ~ dnorm(mu[i],tau)
    mu[i] <- beta[1]+beta[2]*Age[i]+beta[3]*Smoke[i]
    +beta[4]*Age[i]*Smoke[i]
  }
  # The priors on beta and tau are specified here
  beta[1:r] ~ dmvn(beta0[1:r],C0inv[1:r,1:r])
  tau ~ dgamma(a,b)
  # Estimate mean FEV for smokers and nonsmokers aged 10,...,19
  for(i in 1:10){
    meanFEVs[i] <- beta[1]+(beta[2]+beta[4])*(i+9)+beta[3]
    meanFEVns[i] <- beta[1]+beta[2]*(i+9)
  }
  # Easy to estimate relative means and mean differences as well
  RM <- meanFEVns[9]/meanFEVs[9] #RM comparing 18 yo NS to S
  MD <- meanFEVs[9]-meanFEVns[4] #MD comparing 18 yo S to 13 yo NS
  # Predict the FEV for a 20 year-old smoker and nonsmoker
  FEV20s ~ dnorm(mu20s,tau)
  FEV20ns ~ dnorm(mu20ns,tau)
  mu20s <- beta[1]+(beta[2]+beta[4])*20 + beta[3]
  mu20ns <- beta[1]+beta[2]*20
}
```

The WinBUGS code was saved to a file named `FEVWBModel.txt`. The model with a proper reference prior is fitted by setting  $a$  and  $b$  equal to a small value (say, 0.001),  $\beta_0$  equal to an  $r$  vector of all zeros ( $\beta_0 \leftarrow \text{rep}(0, r)$ ), and  $C_0^{-1}$  equal to  $cI_r$  for some small value  $c$  ( $C_0^{-1} \leftarrow \text{diag}(\text{rep}(c, r))$ ). To use a BCJ prior, modifications must be made as illustrated in the R code given next.

To run WinBUGS from within R, prior to executing the R script you must load into R the FEV data and suitably modify the directories used. The WinBUGS code in FEVWBModel.txt is called into the R code using the bugs function, which is part of the R2WinBUGS library. Appendix C contains additional information about importing data into R and the R2WinBUGS library.

```
library(R2WinBUGS) # Brings the appropriate environment into
# the R session
Ytilde <- c(2.8,3,4,3.3) # Specify prior mean vector
D <- diag(c(0.04,0.04,0.04,0.09)) # and cov matrix for mtilde
Xtilde <- matrix(c(1,11,0,0, # Specify Xtilde matrix
                  1,13,1,13,
                  1,16,0,0,
                  1,18,1,18),4,4,byrow=T)
Xtildeinv <- solve(Xtilde) # Invert Xtilde
# Get prior mean vector and precision matrix for beta
beta0 <- c(t(Xtildeinv %*% Ytilde))
C0 <- Xtildeinv %*% D %*% t(Xtildeinv)
C0inv <- solve(C0)
a <- 1.73 # a and b are the hyperparameters of prior on tau
b <- 0.78
n <- length(FEV) # Get the sample size, n
r <- dim(Xtildeinv)[1] # Number of regression coefficients
# Create a list of all the inputs appearing in the WinBUGS code
FEVdataBUGS <- list("n","r","beta0","C0inv",
                     "a","b","FEV","Age","Smoke")
# Identify all objects to be monitored in WinBUGS
parameters <- c("beta","tau","meanFEVs","meanFEVns",
                 "RM","MD","FEV20s","FEV20ns")
# Specify initial values for all stochastic nodes
inits <- list(list(tau=1,beta=c(0,0,0,0),FEV20s=2, FEV20ns=2))
# Call WinBUGS from R to run the model that is in the file
# FEVWBModel.txt. This file is saved in the working.directory,
# which you get to pick. All the inputs needed by WinBUGS are
# in FEVdataBUGS and the initial values are in inits.
FEV.fit <- bugs(FEVdataBUGS, inits, parameters,"FEVWBModel.txt",
                 working.directory="H:\\MyDocuments\\BAYES\\LinReg",
                 n.chains=1, n.iter=60000, n.thin=1, n.burnin=10000)
print(FEV.fit,digits=3)
attach.bugs(FEV.fit)
```

For analyzing other linear regression data with BCJ priors, the user will need to change Ytilde (the vector of elicited means), D (the diagonal matrix containing the elicited variances), Xtilde, as well as the quantities mentioned earlier.

Additional code for fitting linear regression models (e.g., with diffuse priors) is available at the book website. The model with a proper reference prior is fit by deleting Ytilde, D, Xtilde, and Xtildeinv but incorporating the specifications indicated earlier. No changes need to be made to the WinBUGS code.

**EXERCISE 9.13. *FEV Data.*** (a) Analyze the FEV data using the WinBUGS code given earlier with the BCJ prior of Section 9.4.1. (b) For the full FEV data with four predictor variables, construct a partially informative prior based on the prior of Example 9.4.1 by taking the first and third rows of the  $\bar{X}$  matrix given there, the corresponding prior specifications for  $\bar{m}_1$  and  $\bar{m}_3$ , and then using a reference prior on the other coefficients. Run appropriately modified code using this prior. (c) Run

the appropriately modified code on the full FEV data using the usual proper prior approximation to the SIR prior. Give a brief comparison of the results with those from parts (a) and (b). (d) Using the output from (a), create a scatterplot of FEV versus age, and on the same graph, plot mean response versus age for smokers and for nonsmokers.

## 9.5 ANOVA

The methods from Section 5.2 for analyzing one- and two-sample normal data can be extended to compare three or more populations. Fisher's treatment of such data was an *analysis of variance* (ANOVA). Now ANOVA has come to mean the special case of linear models in which only categorical predictor variables are used. One-way ANOVA involves a single multilevel factor; two-way ANOVA has two categorical predictors (and possibly their interaction), etc. As discussed in Subsection 7.4.3, such multilevel factor variables must be transformed into a series of group indicator variables. The simplest analysis of covariance (ACOVA) model has one continuous and one categorical predictor. We discuss two analyses for ANOVA, one that places independent priors on the population means and another that involves a hierarchical prior for the population means.

As a motivating example, recall the Diasorin data of Example 5.2.2. Chronic kidney disease patients with low bone turnover were compared to patients with normal turnover using a two-group normal analysis. In addition, there is a third group of  $n_3 \equiv n_H = 50$  patients with high bone turnover. It is useful to distinguish low bone turnover patients so that they can be properly treated. It is also important to identify patients with high turnover because a different course of treatment is given to them.

The prior information provided by Dr. Herberth for the high bone turnover group was his best guess for the median, 600, with 95% certainty that the median does not exceed 682. Independent prior information for the means of the low and normal groups was discussed in Example 5.2.2 and we continue to use those priors. Similar methods lead us to the prior  $\mu_3 \equiv \mu_H \sim N(6.40, 0.006)$ .

The Diasorin example now provides one-way ANOVA data with the factor "turnover group" having 3 levels: low, normal, and high. The primary scientific issue is whether Diasorin can discern between the three groups. Specifically, the data analysis looks for evidence that  $\mu_1 < \mu_2 < \mu_3$ . We also illustrate how to aid therapeutic decisions by providing the probability of having low, normal, or high turnover conditional on Diasorin score.

First, we present the general method of Bayesian ANOVA and discuss how it relates to linear regression. The importance of assessing equality of variances among groups (*homoscedasticity*) is reviewed along with how homoscedasticity is tied to comparing normal means.

### 9.5.1 Independence Prior

Draw a random sample of size  $n_i$  from population  $i$ ,  $i = 1, 2, 3, \dots, I$ . We assume that population  $i$  has mean  $\mu_i$  and variance  $\sigma^2$ . The variance is the same for each group. The data and corresponding summary statistics are represented below.

Group	Data	Sample Mean	Sample Variance
Population 1	$y_{11}, y_{12}, \dots, y_{1n_1}$	$\bar{y}_1.$	$s_1^2$
Population 2	$y_{21}, y_{22}, \dots, y_{2n_2}$	$\bar{y}_2.$	$s_2^2$
Population 3	$y_{31}, y_{32}, \dots, y_{3n_3}$	$\bar{y}_3.$	$s_3^2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
Population $I$	$y_{I1}, y_{I2}, \dots, y_{In_I}$	$\bar{y}_I.$	$s_I^2$

A common task is to use the data and prior information to determine whether there is statistical evidence that the population means are different, and if so, by how much.

The standard one-way ANOVA model posits

$$y_{ij} | \mu_i, \tau \stackrel{iid}{\sim} N(\mu_i, 1/\tau), \quad i = 1, \dots, I, \quad j = 1, \dots, n_i, \quad (1)$$

where, in addition to homoscedastic variances, we assume independence and normality for all the data. Take independence priors for this model,

$$\mu_i \stackrel{ind}{\sim} N(a_i, 1/b_i) \quad \perp \!\!\! \perp \quad \tau \sim \text{Gamma}(c, d).$$

For small  $I$ , researchers can readily determine  $a_i$  and  $b_i$ . For large  $I$ , a proper reference prior may use  $a_i = 0$ ,  $b_i = 0.000001$ . Often researchers select  $c = d = 0.001$ , mimicking the prior  $p(\tau) = 1/\tau$ .

Reindexing all the observations from  $(i, j)$  to  $k$ , rewrite the ANOVA model (1) as a linear regression

$$y_k = \beta_1 + \beta_2 x_{k2} + \dots + \beta_I x_{kI} + \varepsilon_k, \quad \varepsilon_k | \tau \stackrel{iid}{\sim} N(0, 1/\tau) \quad (2)$$

where  $x_{ki}$  is an indicator variable for whether observation  $k$  belongs to group  $i$ . With an intercept in the model, only  $I - 1$  of the  $x_{ki}$ s are used. The one not used becomes the baseline group (group 1 here) and is chosen by the researcher. There is a one-to-one correspondence between the ANOVA and linear regression parameterizations. The intercept  $\beta_1$  corresponds to the baseline group mean, here  $\mu_1$ , and here, for  $i = 2, \dots, I$ ,  $\beta_1 + \beta_i$  corresponds to  $\mu_i$ . In general  $\beta_i$ ,  $i > 1$ , is the regression coefficient for a group indicator variable and  $\beta_1 + \beta_i$  corresponds to the mean for that group when an intercept is in the model. Alternatively, we can write model (1) as the regression model without an intercept

$$y_k = \mu_1 x_{k1} + \mu_2 x_{k2} + \dots + \mu_I x_{kI} + \varepsilon_k, \quad \varepsilon_k | \tau \stackrel{iid}{\sim} N(0, 1/\tau). \quad (3)$$

Our assumed prior is a BCJ prior for model (3).

Checking the assumptions of normality and equal variances involves plotting histograms, box-plots, residual plots, and normal plots and inspecting the sample variances. These model diagnostics are illustrated in Section 9.6. If one population is much more spread out than the rest, or if any of the histograms were overly skewed, that causes concern over the validity of the sampling model. A standard corrective action transforms the response data to better satisfy the assumptions of ANOVA models, cf. Christensen (1996, Sec. 7.10).

After transforming data to achieve approximate homoscedastic normality, inferences made on the means, which are also the medians, can be reinterpreted on the original measurement scale as inferences for medians. For example, if we conclude that for transformed data  $\mu_1 > \mu_2$ , then for the untransformed data the median response from population 1 exceeds that of population 2 (provided the transformation is increasing). Consequently, more than half of the individuals from population 1 have values exceeding the median of population 2. Unlike means, the median of the transformed data corresponds to the median of the untransformed data. Therefore, we prefer to make inferences about the medians.

As discussed in Subsection 5.2.5, heteroscedasticity is always a serious problem. In ANOVA, just like for two-sample data, it poses no technical problems to Bayesians (unlike frequentists). Posterior inferences are easily obtained by simulation. The real issue remains one of interpretation. Figure 5.3 illustrated that with unequal variances the practical significance of the analysis may involve much more than merely comparing the means of the various populations. Whereas with equal variances, examination of the mean values tells a much larger part of the story. *Again, we find that plotting the predictive distributions is most useful, especially when the data have unequal variances.*

For the heteroscedastic case, let

$$y_{ij} | \mu_i, \tau_i \stackrel{iid}{\sim} N(\mu_i, 1/\tau_i), \quad i = 1, \dots, I, \quad j = 1, \dots, n_i. \quad (4)$$

We again take independence priors for this model, now

$$\mu_i \stackrel{ind}{\sim} N(a_i, 1/b_i) \quad \perp \!\!\! \perp \quad \tau_i \stackrel{ind}{\sim} \text{Gamma}(c_i, d_i).$$

Table 9.3: Posterior summaries from ANOVA of the Diasorin data.

Parameter	(a) Informative Prior on $\mu_i$ s				
	Mean	sd	2.5%	Median	97.5%
$\mu_L$	4.86	0.05	4.76	4.86	4.96
$\mu_N$	5.40	0.05	5.29	5.40	5.50
$\mu_H$	6.26	0.07	6.12	6.26	6.39
$\tau$	1.20	0.19	0.85	1.19	1.60
	(b) Large Variance for $\mu_i$ s				
	Mean	sd	2.5%	Median	97.5%
$\mu_L$	4.70	0.20	4.31	4.71	5.10
$\mu_N$	5.49	0.23	5.05	5.49	5.94
$\mu_H$	5.85	0.12	5.61	5.85	6.10
$\tau$	1.32	0.21	0.95	1.31	1.76
	(c) Approximate SIR prior				
	Mean	sd	2.5%	Median	97.5%
$\mu_L$	4.70	0.20	4.31	4.71	5.10
$\mu_N$	5.49	0.23	5.05	5.49	5.94
$\mu_H$	5.85	0.12	5.61	5.85	6.10
$\tau$	1.32	0.21	0.95	1.31	1.76

As before, for small  $I$  researchers can readily determine  $a_i$  and  $b_i$  and for large  $I$  a proper reference prior may be used. Again, researchers often select  $c_i = d_i = 0.001$  mimicking the prior  $p(\tau_i) = 1/\tau_i$ .

Models (2) and (3), the regression forms of model (1), share our standard assumption of homoscedasticity for regression models. The heteroscedastic version of model (2) specifies

$$y_k = \beta_1 + \beta_2 x_{k2} + \cdots + \beta_I x_{kI} + \varepsilon_k, \quad \varepsilon_k \stackrel{ind}{\sim} N(0, 1/\tilde{\tau}_k)$$

with

$$\tilde{\tau}_k \equiv \tau_1 x_{k1} + \tau_2 x_{k2} + \cdots + \tau_I x_{kI}.$$

Heteroscedastic regression models are much more difficult for frequentists to analyze.

**EXAMPLE 9.5.1.** *Bayesian ANOVA of the Diasorin Data.* The residuals from a least squares (SIR) fit suggest that a normal sampling model is viable for the log-transformed Diasorin data. Using the informative priors on the means and  $\text{Gamma}(0.001, 0.001)$  priors on all precisions, the heteroscedastic model (4) has  $\text{DIC} = 229.4$  but the equal variance model (1) is preferred since its  $\text{DIC}$  is smaller at 225.7. (DIC as a model selection criterion was discussed in Section 4.9.)

First consider analyzing and interpreting the data on the log scale, that is, we simply regard the transformed data as the data. Some posterior results based on three priors are presented in Table 9.3. In each prior a  $\text{Gamma}(0.001, 0.001)$  was used for the common precision. The first prior (a) used the informative normal distributions for the means; the second prior (b) used normal distributions that were centered as in (a) but with variance 1000; the third prior (c) used proper diffuse priors for all means to approximate the SIR prior.

Results based on priors (b) and (c) are virtually identical to each other, and qualitatively similar to results based on prior (a). Estimated medians increase from low to normal to high in all analyses, but the estimated median for the high group is considerably larger under prior (a). All of the posterior standard deviations are appreciably smaller under prior (a) than the others.

In addition to the results in Table 9.3, under all priors we are at least 95% certain that the median in the normal group is larger than the median in the low group, and that high is also bigger than low. With priors (b) and (c), but not (a), the 95% probability interval for  $\mu_H - \mu_N$  covers 0.

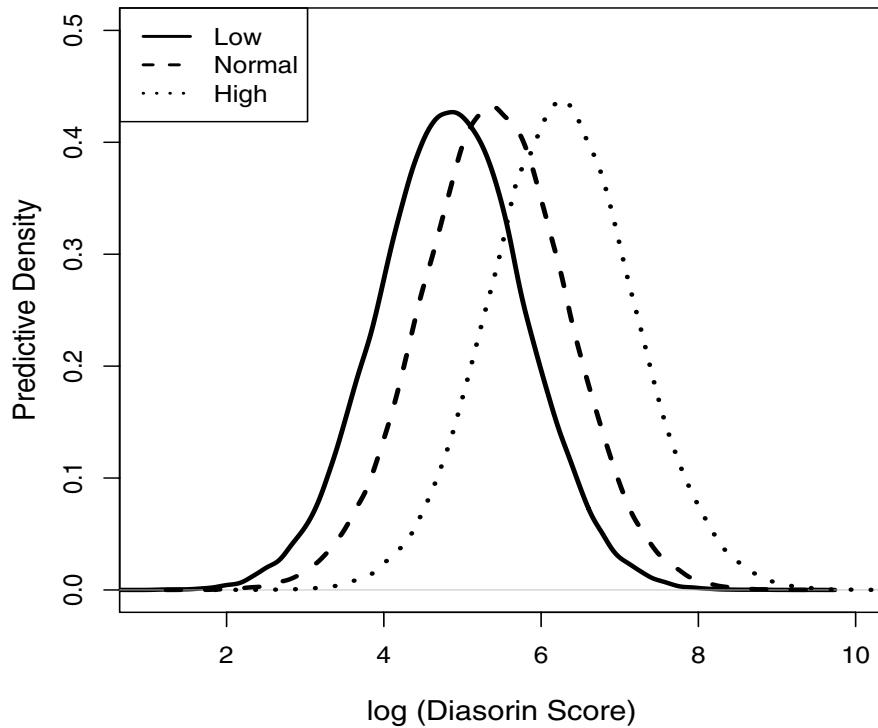


Figure 9.2: *Predictive densities for log-Diasorin scores.*

Under prior (c), the posterior probability that  $\mu_H$  is larger than  $\mu_N$  is 0.92. The posterior results are consistent for the low and normal populations with those in Table 5.3. The posterior probability  $P(\mu_1 < \mu_2 < \mu_3 | Y)$  equals 1 under prior (a), and equals 0.91 under priors (b) and (c).

Figure 9.2 shows predictive distributions from prior (a) for all three groups on the log scale. While the distributions are clearly distinct, they show substantial overlap.

The diagnostic potential of Diasorin to distinguish among bone turnover groups is supported by these results, although the predictive distributions suggest that Diasorin is far from a perfect discriminator. Before a definitive conclusion on Diasorin's clinical relevance is reached, a follow-up large-scale study may be warranted that includes various demographic and health variables.

When transforming the data, the analysis can be conducted on the transformed scale, but often answers are desired on the original scale. In the case of Diasorin, it is unclear whether this is an issue, but if we were dealing with a familiar measurement like blood pressure, clearly interpretations should be presented on the original scale of the data. Prior information should always be elicited on the scale most familiar to the expert.

Table 9.4 gives posterior results for priors (a) and (c) similar to those in Table 9.3 but on the original scale. The parameters are the medians  $e^{\mu_i}$  and the relative medians  $e^{\mu_i - \mu_j}$  for comparing low, normal, and high groups. Using our informative prior (a), just as the 95% probability intervals for  $\mu_i - \mu_j$  all excluded 0, the intervals for the relative medians all exclude 1. Thus for any two medians we are more than 97.5% sure that they differ. More interestingly (since we already knew that), we are 97.5% sure that the median in the high group is at least 3.41 times that for the low group.

Table 9.4: Posterior summaries from ANOVA of the Diasorin data; original scale.

(a) Informative Prior on $\mu$					
Inference	Mean	sd	2.5%	Median	97.5%
Med <sub>L</sub>	129.2	6.72	116.5	129.0	142.8
Med <sub>N</sub>	220.5	11.36	199.1	220.2	243.7
Med <sub>H</sub>	522.0	36.03	455.5	520.5	596.8
Med <sub>N</sub> /Med <sub>L</sub>	1.71	0.13	1.48	1.71	1.97
Med <sub>H</sub> /Med <sub>L</sub>	4.05	0.35	3.41	4.04	4.78
Med <sub>H</sub> /Med <sub>N</sub>	2.37	0.20	2.00	2.36	2.80
(c) Approximate SIR Prior					
	Mean	sd	2.5%	Median	97.5%
Med <sub>L</sub>	112.7	22.98	74.32	110.50	164.10
Med <sub>N</sub>	248.40	57.12	155.30	242.10	377.90
Med <sub>H</sub>	349.80	43.84	272.10	347.10	444.10
Med <sub>N</sub> /Med <sub>L</sub>	2.30	0.72	1.21	2.19	3.99
Med <sub>H</sub> /Med <sub>L</sub>	3.23	0.78	1.97	3.14	4.99
Med <sub>H</sub> /Med <sub>N</sub>	1.48	0.39	0.86	1.43	2.38

The estimated group medians are 129.0, 220.2, and 520.5 for low, normal, and high, respectively. The analysis for prior (c) is similar.

All of our results for prior (a), along with other results that we did not report, were obtained using the following WinBUGS code.

```
# Model data as log normal
model{ # First part is for analysis on log scale
  for(i in 1:n[1]) {low[i]~dlnorm(mu[1],tau)}
  for(i in 1:n[2]) {normal[i]~dlnorm(mu[2],tau)}
  for(i in 1:n[3]) {high[i]~dlnorm(mu[3],tau)}
# Prior specification
  mu[1]~dnorm(4.87,347.12)
  mu[2]~dnorm(5.39,357.42)
  mu[3]~dnorm(6.40,166.67)
  tau~dgamma(0.001,0.001)
# Test for ordering of medians
  P <- step(mu[3]-mu[2])*step(mu[2]-mu[1])
# Predictive densities for three future (log-scale) observations
  lowf.ls~dnorm(mu[1],tau)
  normalf.ls~dnorm(mu[2],tau)
  highf.ls ~ dnorm(mu[3],tau)
# Additional inferential objects
  diff21 <- mu[2]-mu[1]
  diff31 <- mu[3]-mu[1]
  diff32 <- mu[3]-mu[2]
  prob21 <- step(diff21)
  prob31 <- step(diff31)
  prob32 <- step(diff32)
# The following part is for analysis on original data scale
  med[1] <- exp(mu[1])
  med[2] <- exp(mu[2])
```

```

med[3] <- exp(mu[3])
relmed21 <- med[2]/med[1]
relmed31 <- med[3]/med[1]
relmed32 <- med[3]/med[2]
# Simulate predictive densities on untransformed scale
lowf <- exp(lowf.ls)
normalf <- exp(normalf.ls)
highf <- exp(highf.ls)
} # List of inputs follows
list(n=c(19,15,50),
low=c(91,46,95,60,33,410,105,43,189,1097,54,178,114,137,
233,101,25,70,357),
normal=c(370,267,99,157,75,1281,48,298,268,62,804,
430,171,694,404),
high=c(75,52,1378,555,331,231,472,263,120,46,650,349,251,
492,759,96,627,171,1584,69,368,509,486,354,351,
839,88,162,1041,383,234,1130,503,244,606,457,460,
283,767,576,628,239,583,428,452,723,201,406,422,243))
# Initial values
list(mu=c(4.87,5.39,6.4),tau=1,
lowf.ls=100,normalf.ls=200,highf.ls=300)

```

The use of the log transformation in WinBUGS is facilitated by the existence of code for log-normal distributions. Using, say, a square root transformation would require coding the square root and using normal distributions.

In the Diasorin example,  $1 = P(\mu_1 \leq \mu_2 \leq \mu_3 \mid Y) = E[I_{[0,\infty)}(\mu_3 - \mu_2)I_{[0,\infty)}(\mu_2 - \mu_1) \mid Y]$  under prior (a). With continuous priors, there is zero probability for any two means being equal so the probability that they are different is 1. More importantly, we might want the probability that they are substantially different. Borrowing a trick from frequentist ANOVA, we could compute the sample variance of the  $\mu_i$ s, say,  $s_\mu^2$  and find the probability that  $s_\mu > \varepsilon$  for some positive number  $\varepsilon$ . Here “substantially different” is defined implicitly by  $\varepsilon$ . Bayesian analysis provides great flexibility to define and examine functions of the parameters that are relevant to the problem at hand.

**EXERCISE 9.14.** Using the informative prior (a) for the Diasorin data, write WinBUGS code to verify the estimates in Tables 9.3 and 9.4. Give pictures of the three predictive densities under priors (a) and (c) both on the transformed scale, as in Figure 9.2, and on the untransformed scale.

**EXERCISE 9.15.** Using the informative prior (a) for the Diasorin data, run the WinBUGS code and obtain the DIC statistic. Then modify the code to handle unequal variances/precisions, with the same prior for the means but with independent Gamma(0.001, 0.001) priors for all precisions. Monitor the ratios of standard deviations comparing the first and second, first and third, and second and third groups. Obtain the DIC statistic and compare it to the DIC for homogeneous variances. Thus verify our previous conclusion that equal variances are reasonable. Nonetheless, explain how inferences would change if you assumed heterogeneous variances. Should you care about the changes? See Subsection 4.9.3 for discussion of DIC.

### 9.5.1.1 Allocation and Diagnosis

For the Diasorin data, the practicing nephrologist is interested in diagnosing patients with bone turnover abnormalities (low or high) and recommending treatment accordingly. With such small sample sizes and only one predictor variable we have little hope of constructing a definitive classification rule. Instead we illustrate a statistical approach that provides information useful to a

Table 9.5: Posterior medians and 95% PIs for group-probabilities.

$y_f$	$\Pr[\text{low}   y_f, Y]$	$\Pr[\text{normal}   y_f, Y]$	$\Pr[\text{high}   y_f, Y]$
log(50)	0.77 (0.70, 0.84)	0.18 (0.13, 0.22)	0.05 (0.02, 0.10)
log(150)	0.53 (0.50, 0.57)	0.25 (0.23, 0.26)	0.22 (0.17, 0.26)
log(250)	0.39 (0.36, 0.41)	0.25 (0.23, 0.27)	0.36 (0.33, 0.39)
log(350)	0.29 (0.24, 0.33)	0.23 (0.22, 0.25)	0.48 (0.45, 0.52)
log(450)	0.23 (0.17, 0.28)	0.21 (0.19, 0.23)	0.56 (0.51, 0.62)
log(550)	0.18 (0.12, 0.24)	0.19 (0.17, 0.21)	0.63 (0.56, 0.70)
log(650)	0.15 (0.09, 0.21)	0.17 (0.15, 0.20)	0.68 (0.61, 0.75)
log(750)	0.12 (0.07, 0.19)	0.16 (0.13, 0.19)	0.72 (0.64, 0.79)
log(850)	0.10 (0.06, 0.17)	0.15 (0.11, 0.18)	0.75 (0.67, 0.83)

diagnostician, namely the probability that a patient has low, normal, or high bone turnover given their Diasorin score.

Having analyzed our one-way ANOVA data  $Y$ , suppose we are confronted with a new observation  $y_f$  without knowing its group. What can we say about the group? In this context, model (1) defines  $I$  potential models for the data. We do not know the parameters in model (1), so our models are actually the predictive distributions

$$f_i(y_f | Y) \equiv \int f(y_f | \mu_i, \tau) p(\mu_i, \tau | Y) d\mu_i d\tau.$$

Suppose we also have prior probabilities for each group, say,  $\pi_1, \dots, \pi_I$ . By Bayes' Theorem

$$\Pr[\text{Group } r | y_f, Y] = f_r(y_f | Y) \pi_r / \sum_{i=1}^I f_i(y_f | Y) \pi_i.$$

For the Diasorin data

$$\begin{aligned} \Pr(\text{low} | y_f, Y) &= \frac{f_1(y_f | Y) \pi_1}{f_1(y_f | Y) \pi_1 + f_2(y_f | Y) \pi_2 + f_3(y_f | Y) \pi_3} \\ \Pr(\text{normal} | y_f, Y) &= \frac{f_2(y_f | Y) \pi_2}{f_1(y_f | Y) \pi_1 + f_2(y_f | Y) \pi_2 + f_3(y_f | Y) \pi_3} \\ \Pr(\text{high} | y_f, Y) &= \frac{f_3(y_f | Y) \pi_3}{f_1(y_f | Y) \pi_1 + f_2(y_f | Y) \pi_2 + f_3(y_f | Y) \pi_3} \end{aligned}$$

where now  $\pi_i$  is the population prevalence among stage 5 chronic kidney disease patients of low ( $i = 1$ ), normal ( $i = 2$ ), and high ( $i = 3$ ) bone turnover. We used values for the  $\pi_i$ s obtained from Dr. Herberth ( $\pi_1 = 0.40$ ,  $\pi_2 = 0.20$ ,  $\pi_3 = 0.40$ ).

Table 9.5 and Figure 9.3 give posterior estimates of these probabilities for several values of  $y_f$ . Among chronic kidney disease patients with  $y_f = \log(50)$ , we estimate that 77% have low turnover and only 5% have high turnover. Among patients with  $y_f \geq \log(550)$ , more than 60% have high turnover and fewer than 20% have low turnover. Particularly, for  $y_f = \log(850)$  the probability of having high turnover is 0.75 and the probability of having low turnover is only 0.10. Note that with these prior group probabilities, the normal group is never the group with the highest probability, regardless of the Diasorin score.

In this process, it does not matter if you use the predictive distributions for the Diasorin scores or those for the log-Diasorin scores. In Exercise 4.8 we establish that transformed data values yield proportional likelihood functions, so the posterior distribution of the parameters is the same regardless of any data transformation. Similarly, for the purpose of this calculation, it does not matter

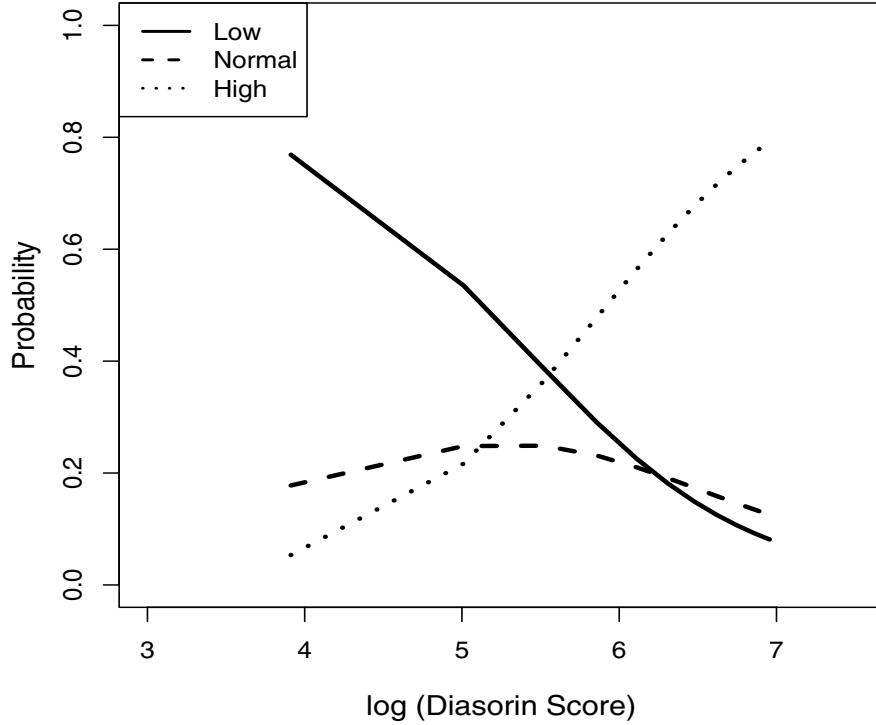


Figure 9.3: Probability of group membership versus log-Diasorin score.

if we use the predictive distributions for the original data or the transformed data. From Proposition B.4 if we transform  $w = G(y)$  and apply it to each  $f_i(y|Y)$ , we get new densities  $g_i(w|Y)$  with  $f_i(y_f|Y) \propto g_i(w_f|Y)$  for  $w_f = G(y_f)$  where the constant of proportionality depends on  $y_f$  (or  $w_f$ ) but does not depend on  $i$ , so it cancels out in Bayes' Theorem.

This application of Bayes' Theorem is simple and intuitive, but it assumes that  $y_f$  comes from a distribution  $f_i(y|Y)$  rather than  $f(y|\mu_i, \tau)$ . The procedure ignores the information that  $y_f$  contains about the parameters. For example, if  $y_f$  is larger than any observation in  $Y$ , it would have some effect in raising the estimates of the  $\mu_i$ s. We feel that the simplicity of the current procedure outweighs the loss of information. Note that the procedure applies to the heteroscedastic model (4) just as well as to model (1).

To obtain this allocation analysis from the WinBUGS code given previously, include

```
for(i in 1:11){
  aL[i] <- piL*sqrt(tau/(2*3.14159))*(1/x[i])
  *exp(-(tau/2)*(log(x[i])-mu[1])*(log(x[i])-mu[1]))
  bL[i] <- piN*sqrt(tau/(2*3.14159))*(1/x[i])
  *exp(-(tau/2)*(log(x[i])-mu[2])*(log(x[i])-mu[2]))
  cL[i] <- piH*sqrt(tau/(2*3.14159))*(1/x[i])
  *exp(-(tau/2)*(log(x[i])-mu[3])*(log(x[i])-mu[3]))
  pL[i] <- aL[i]/(aL[i] + bL[i] + cL[i])
  pN[i] <- bL[i]/(aL[i] + bL[i] + cL[i])
  pH[i] <- cL[i]/(aL[i] + bL[i] + cL[i])
}
```

in the model statement. The posterior means of  $pL[i]$ ,  $pN[i]$ , and  $pH[i]$  are the estimated probabilities. Also add to the list of inputs

```
x <- c(50, 150, 250, 350, 450, 550, 650, 750, 850, 950, 1050)
```

In either the hetero- or homoscedastic cases, when  $I$  is moderate to large it is difficult to specify subjective priors for all the  $\mu_i$ s. As an alternative to using reference priors, one might model the  $\mu_i$ s exchangeably, with  $\mu_i | \theta, \tau_\mu \stackrel{iid}{\sim} N(\theta, \tau_\mu)$ . Manageable prior information can be introduced by placing an informative prior on  $(\theta, \tau_\mu)$ . This hierarchical prior is discussed in the next subsection.

### 9.5.2 Hierarchical Priors and Models

With the prior that we have assumed, if the scientist were told the value of one  $\mu_i$ , the priors for the other means would not change. An alternative scenario is that knowing one  $\mu_i$  makes the scientist think that other means should take similar values. In such cases specifying real prior information for each group would not be easy, especially if there are more than a few groups. We now incorporate this scenario by using a hierarchical (two-stage or multi-level) prior as discussed in Section 4.12.

Alternatively, it may be the case that the  $\mu_i$ s can be regarded as having come from their own distribution. This would be especially true if the number of groups  $I$  is moderately large and if the groups may be regarded as exchangeable. Such a distribution would have its own parameters. In this case, the distribution of  $\mu_i$ s is part of the model for the data, and the parameters of that distribution would be model parameters that would require specification of a prior. As mentioned in Section 4.12, the two approaches to modeling are structurally identical.

Consider the homoscedastic case and assume that the scientist regards the population means as exchangeable. This implies that the marginal distributions for each mean are identical and allows the means to be dependent, see Section 4.2. Our Bayesian model is specified

$$y_{ij} | \mu_i, \tau \stackrel{ind}{\sim} N(\mu_i, 1/\tau), \quad (5)$$

$$\mu_i | \theta, \tau_\mu \stackrel{iid}{\sim} N(\theta, 1/\tau_\mu), \quad (6)$$

$$p(\theta, \tau, \tau_\mu) = p(\theta)p(\tau)p(\tau_\mu). \quad (7)$$

In the case of specifying a hierarchical prior, the sampling distribution is specified by (5) and the prior by (6) and (7). In the case of specifying a hierarchical model for the data, the sampling distribution of the data is determined by (5) and (6) and the prior by (7). Often a normal prior is specified for  $\theta$  with Gamma or Uniform priors for  $\tau$  and  $\tau_\mu$ .

The prior on  $\theta$  is easy to obtain since  $\theta = E(\mu_i | \theta, \tau_\mu)$  for all  $i$ . We ask our expert to think about the average for the population of means. Specifically, we elicit a best guess and a tentative upper bound. As we have done so often, we use these to identify the mean and precision for  $\theta \sim N(a, 1/b)$ .

The prior on the error precision  $\tau$  is Gamma( $c, d$ ). For a typical population, say  $i$ , with supposedly known mean  $\mu_i$ , we elicit information about, say, the 90th percentile of the observations. Then proceed as we have repeatedly illustrated. If the scientist believes the sampling model, the actual choice of  $i$  and  $\mu_i$  are irrelevant.

The prior for  $\tau_\mu$  is obtained as in Section 8.5 for a binomial mixed model. See Exercise 9.17 for details.

A proper reference prior uses independence with  $\text{Gamma}(0.001, 0.001)$  distributions for  $\tau$  and  $\tau_\mu$  and a  $N(0, 10^6)$  for  $\theta$ . A reasonable theoretical reference prior for  $(\theta, \tau)$  is  $p(\theta, \tau) \propto 1/\tau$ . It is tempting to take  $\tau_\mu$  independent with  $p(\tau_\mu) \propto 1/\tau_\mu$ . Unfortunately, it is well known that this leads to an improper posterior for  $\tau_\mu$ , cf. Gelman et al. (2004, Problem 5.8). This is not an issue for proper priors. The impropriety of the marginal posterior for  $\tau_\mu$  is intuitively due to a lack of information in the data for  $\tau_\mu$ , especially when  $I$  is small. We are trying to estimate the variability in the  $\mu_i$ s when the  $\mu_i$ s are not observed. Even if they were observed, it would be difficult to estimate the variance/precision with small  $I$ .

The choice of an informative Gamma prior can also be problematic if it is not well conceived. With small  $I$ , the posterior is likely to be similar to the prior. It is probably best with small  $I$  to use independent priors on the  $\mu_i$ s without additional hierarchical structure.

When specifying a hierarchical prior,  $\theta$  and  $\tau_\mu$  are part of the prior, so there would be little interest in them. However, if the sampling model is regarded as hierarchical, they are key parameters for the data and much interest is focussed on them. We go into greater detail on this distinction in Chapter 10. Here our interest remains on the  $\mu_i$ s and their differences, just as with the independence prior.

A WinBUGS template is presented below. In our analysis of model (1), we changed from subscripts  $ij$  to a single subscript  $k$  in models (2) and (3). Our WinBUGS code used a single column of data. Now, the matrix  $y[,]$  contains the response data. It must be a complete rectangular array. That only occurs when the sample sizes are the same for each group. For unequal sample sizes, like the Diasorin data, the response matrix can be completed by appropriately filling in NAs. Alternatively, the contribution to the likelihood from the data for each group can be specified using separate “for” loops in WinBUGS.

```
model{
  for(i in 1:I){
    mu[i]~dnorm(theta,tau.mu)
    for(j in 1:n){
      y[i,j]~dnorm(mu[i],tau)
    }
  }
  tau~dgamma(c,d)
  theta~dnorm(a,b)
  tau.mu~dgamma(c.mu,d.mu)
  sigma <- sqrt(1/tau)
}
```

**EXERCISE 9.16. *Diasorin Data.*** Analyze the Diasorin data using the ANOVA model with a diffuse but *proper* hierarchical prior on the  $\mu_i$ s. Perform a sensitivity analysis by varying the Gamma prior on  $\tau_\mu$ . Compare your analysis with the analysis in Example 9.5.1.

**EXERCISE 9.17.** Construct an informative Gamma prior for  $\sigma_\mu = 1/\sqrt{\tau_\mu}$  using the following information. Assume that  $\mu_i$  iid  $N(\theta, \sigma_\mu^2)$ . Let  $\gamma_{0.9} = \theta + 1.28\sigma_\mu$  be the 90th percentile of the distribution of means. Let  $\gamma_{0.9,0}$  be the expert’s best guess for  $\gamma_{0.9}$  and let  $\theta_0$  be their best guess for  $\theta$ . Finally, the expert is 95% sure that  $\gamma_{0.9}$  is less than  $\tilde{\gamma}$ , conditional on all of the other information. Assuming that  $\theta$  is independent of  $\sigma_\mu$  (a) derive a suitable Gamma prior for  $\sigma_\mu$  based on this information and (b) modify the argument to derive a suitable Gamma prior for  $\tau_\mu$ .

## 9.6 Model Diagnostics

Good statisticians question whether their model is an adequate approximation to reality. For two or more well-defined alternative models, Bayesian methodology can check which fits the data best. But Bayesian statistics is not well suited to asking the more nebulous question of whether anything could be wrong with the model. As discussed in Section 4.1, significance testing is designed to address exactly that question.

Diagnostics are tools used to check whether the model assumptions look reasonable. Formally, significance testing only addresses the issue of whether there is evidence that something is wrong. When something is wrong, the nature of the test statistic may give hints as to what is wrong. For example, a test statistic that focusses on “homoscedasticity” might well turn out significant because of heteroscedasticity, but it might also show significance because of nonnormality. Based on the test alone we cannot know which. The test itself does not suggest a specific alternative model to

accommodate either heteroscedasticity or nonnormality. Knowing that there is a problem forces us to look deeper and find more appropriate models.

Standard linear regression models assume independence, linearity, normality, and constant variance. There are a variety of methods available to check these assumptions based on least squares estimates, cf. Christensen (2002; 2005 Chapter 13). These include looking at an initial scatterplot matrix to determine marginal relationships between the response and continuous, ordinal, or nominal predictor variables, residual plots, and tests for nonconstant variance and lack of fit. Independence is usually verified on the basis of sampling design, but a plot of residuals versus time-order may be appropriate.

In the regression setting with  $\text{Cov}(Y | \beta, \tau) = (1/\tau)I_n$ , residuals are estimates of the error terms  $\varepsilon_i = y_i - x'_i\beta$ . For evaluating the *sampling distribution* of the data, they would be estimated from the data alone. Based on least squares estimates, these residuals, as well as the Studentized/standardized and deleted residuals discussed later, are available from virtually any statistical package. Estimating the residuals from the posterior as  $y_i - x'_i\hat{\beta}$  with  $\hat{\beta} = E(\beta | Y)$  serves to incorporate prior information into the subsequent evaluations although with the SIR prior,  $\hat{\beta}$  is the least squares estimate.

Using least squares, the Studentized (also called standardized) residuals are residuals that are divided by their *frequentist* estimated standard deviations. The *Studentized residual* is

$$S_i \equiv \frac{y_i - x'_i\hat{\beta}}{\sqrt{MSE(1-h_i)}},$$

where  $h_i = x'_i(X'X)^{-1}x_i$ . The measure  $h_i$ , called the “leverage,” takes on values between 0 and 1 and increases with the distance between  $x_i$  and  $\bar{x}$ , the average of the vectors  $x_j$  in the data (Christensen, 2002, Section 13.1). If  $x_i$  is far removed from the average, leading to  $h_i \doteq 1$ , the regression surface is pulled or leveraged towards the  $i$ th observation,  $y_i$ , since the frequentist standard error of  $(y_i - x'_i\hat{\beta})$  will be near zero.

*Deleted residuals* have the form

$$T_i \equiv \frac{y_i - x'_i\hat{\beta}_{(i)}}{\sqrt{MSE_{(i)}[1 + x'_i(X'_{(i)}X_{(i)})^{-1}x_i]}},$$

where  $MSE_{(i)}$  and the least squares estimate  $\hat{\beta}_{(i)}$  are computed with case  $i$  having been removed from the data, and similarly  $X_{(i)}$  is  $X$  with case (row)  $i$  removed. A significance test compares  $T_i$  to a  $t(n-r-1)$  distribution. These residuals show up in Bayesian calculations with the SIR prior (cf., Johnson and Geisser, 1983), so there is also a Bayesian justification for their use that we discuss shortly.

As discussed in Section 7.3, if we have the correct regression model, residuals when plotted against any variable should look like a band of noise with no discernable pattern. In particular, any fitted trend line should be horizontal at zero. This is true for variables that are in the model and for variables left out of the model. The most commonly used variable in such plots consists of the fitted values  $\hat{y}_i = x'_i\hat{\beta}$ .

The  $\varepsilon_i$ s are assumed to be normal. This can be checked by examining a normal probability (rankit) plot of the residuals.

**EXAMPLE 9.6.1. FEV Data.** For the linear regression model with

$$E[\text{FEV} | A, S] = \beta_1 + \beta_2 A + \beta_3 S + \beta_4 A^*S,$$

where  $A = \text{Age}$  and  $S = \text{Smoke}$ , the least squares estimates and the posterior mean of  $\beta$  using proper reference priors are both approximately  $\hat{\beta} = (0.67, 0.21, 1.72, -0.14)'$ . Rankit and residual plots are presented separately for smokers and nonsmokers in Figure 9.4. The nonparametric trend lines in the residual plots help discern any deviation from a horizontal line at zero. They were computed using

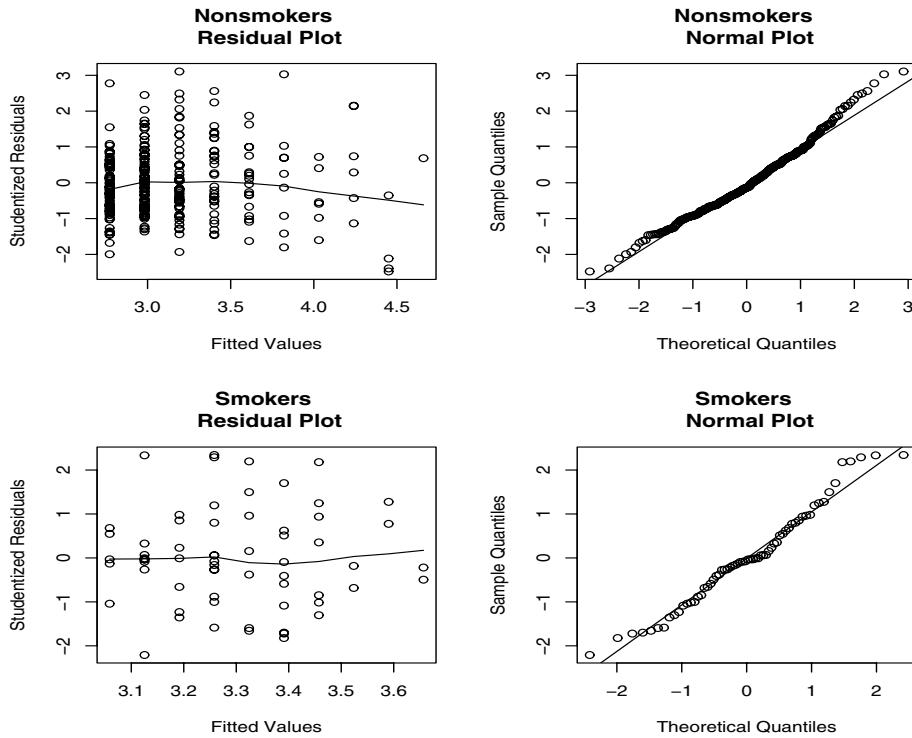


Figure 9.4: *FEV diagnostic plots for linear trends in age.*

*LOWESS* (local weighted scatterplot smoothing), see Cook and Weisberg (1999, Section 9.5.2). The normal plots suggest a slight right skew in these distributions, more so for nonsmokers. (These are consistent with histograms of the residuals.) The constant variance condition may also be in question for older kids (higher fitted values); however, data are sparse for these kids.

Interpreting these plots is obviously subjective. One might conclude that there is no definitive indication of departure from normality or nonconstant variance across age based on these graphics. A log or other transformation could be applied if normality were in question. In our experience, normal plots frequently deviate from the straight line in the tails.

The LOWESS curve in the residual plot for nonsmokers is slightly bowed. A similar shaped LOWESS curve occurs on a plot (not shown) of the Studentized residuals versus age. We therefore investigated the model that includes a separate quadratic trend in age for smokers and nonsmokers:

$$E[\text{FEV} | \text{A}, \text{S}] = \beta_1 + \beta_2 \text{A} + \beta_3 \text{A}^2 + \beta_4 \text{S} + \beta_5 \text{A} * \text{S} + \beta_6 \text{A}^2 * \text{S}.$$

A biological argument in support of this model is that FEV does not increase without bound. It may be expected to plateau and the curvature of a parabola may adequately capture this trend over the age range considered here. (One would certainly not want to extrapolate the quadratic trend very far beyond the observed data.)

The Studentized residuals from the least squares fit of this new model, as plotted in Figure 9.5, suggest that the quadratic trend may have improved the fit for nonsmokers. Additionally, the spread of the residuals appears similar across fitted values, supporting the assumption of equal variance. The rankit plots (not shown) resemble those in Figure 9.4.

**EXERCISE 9.18.** (a) Run linear regression models in WinBUGS with proper reference priors for the FEV data (i) with only the variables Smoke and Age, (ii) with Smoke, Age, and a Smoke by

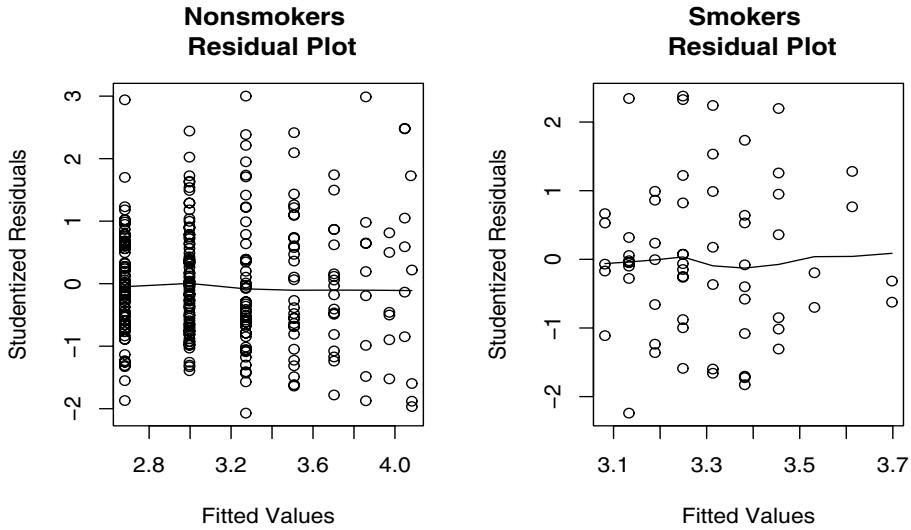


Figure 9.5: Diagnostic plots with a quadratic trend in age.

Age interaction, (iii) with separate quadratic trends in age for smokers and for nonsmokers, and (iv) with a quadratic for nonsmokers and a separate linear trend for smokers. In each case, obtain the DIC statistic. (b) From the model with the smallest DIC statistic, obtain posterior inferences for the regression coefficients and obtain the posterior probability that each regression coefficient is positive. What does this analysis suggest about the associations between FEV and age for smokers and nonsmokers? It may help to graph curves of estimated mean response against age for smokers and nonsmokers on the same plot.

Potential outliers can also be spotted from residual plots. Data points with Studentized residuals that are large in magnitude (say  $> 3$ ) are not being fitted well by the model, and may be considered outliers. Using the iterated expectation formula and the SIR prior,

$$\begin{aligned} E[(y_i - x'_i \beta)^2 | Y] &= E\left\{[(y_i - x'_i \hat{\beta}) + (x'_i \hat{\beta} - x'_i \beta)]^2 | Y\right\} \\ &= (y_i - x'_i \hat{\beta})^2 + E_{\sigma^2|Y}\{E[\text{Var}(x'_i \beta) | \sigma^2, Y]\} \\ &= (y_i - x'_i \hat{\beta})^2 + E(\sigma^2 | Y)h_i, \end{aligned}$$

where the cross product term is  $2(y_i - x'_i \hat{\beta})E[(x'_i \hat{\beta} - x'_i \beta) | Y] = 0$ . Thus, the quality of our predictions depends both on the residual and the leverage. For example, if we had two cases in the data with the same value for  $y_i - x'_i \hat{\beta}$ , and if one of them had high leverage and the other low, the high leverage case would tend not to be close to the “true” regression surface as often as the one with smaller leverage.

Johnson and Geisser (1983, 1985) developed Bayesian methods for detecting the influence that individual cases in the data have on the prediction of future observations and on estimating parameters. While a number of measures were considered there, the simplest one turns out to be proportional to Cook’s distance measure (Cook, 1977),

$$D_i = S_i^2 \frac{h_i}{1 - h_i}.$$

This is the square of the Studentized residual times a monotone function of the leverage.

The simplest interpretation of  $D_i$  is as proportional to the change between the predictions based on full and case deleted data. Technically, it is proportional to the squared Euclidean distance  $(X\hat{\beta} - X\hat{\beta}_{(i)})'(X\hat{\beta} - X\hat{\beta}_{(i)})$ . If we were to predict a future vector of cases with the same covariates seen in the current data using the SIR prior, the mean vector of that predictive distribution would be  $X\hat{\beta}$ . If we obtained the predictive distribution of the same future vector, only based on data with case  $i$  removed, the mean vector would be  $X\hat{\beta}_{(i)}$ .  $\hat{\beta}$  is the posterior mean of  $\beta$  based on the entire data set, while  $\hat{\beta}_{(i)}$  is the posterior mean based on the data with case  $i$  removed. The distance between these two vectors of parameter estimates, as based on an appropriate sense of distance determined by  $[(dfe - 2)/dfe]\text{Cov}(\beta|Y)^{-1} = (X'X)\text{MSE}^{-1}$ , gives the same result as comparing predictions. Thus Cook's distance is a measure of the collective effect of deleting a case either on the prediction of data that are just like the observed data, or on the estimation of  $\beta$ .

To develop a "Bayesian" significance test for an outlier, consider a future case that will have predictor information  $x_f = x_i$  and which will be predicted using the data  $Y_{(i)}$ . Then

$$T_{fi} \equiv \frac{y_f - x_i'\hat{\beta}_{(i)}}{\sqrt{\text{MSE}_{(i)}[1 + x_i'(X'_{(i)}X_{(i)})^{-1}x_i]}} \Big| x_i, Y_{(i)} \sim t(n - r - 1).$$

We can use this distribution to evaluate whether the observed value of  $T_i$  defined earlier is consistent with the model by calculating the  $p$ -value  $\Pr(|T_{fi}| > |T_i| | x_i, Y_{(i)})$ . Case  $i$  is rejected at level  $\alpha$  if  $|T_i| > t(1 - \alpha/2, n - r - 1)$ . It can be seen in Johnson and Geisser (1983) that

$$T_i^2 = \frac{(n - r - 1)S_i^2}{n - r - S_i^2},$$

so it is easy to obtain  $T_i^2$  after obtaining the Studentized residuals  $S_i$ .

Typically, one scans through the entire data to identify  $\max_i |T_i|$  and uses it to test whether outliers exist. To perform a test, one really needs the distribution of  $\max_i |T_{fi}|$ , which is difficult to find. Using Bonferroni's inequality (finite subadditivity), it can be shown that the  $p$ -value from the correct test may be as large as  $n$  times the  $p$ -value computed by comparing the observed value  $\max_i T_i$  to a  $t(n - r - 1)$  distribution. Thus if  $n = 100$ , you would need the  $t(n - r - 1)$  test  $p$ -value to be less than 0.05/100 to be sure that the correct  $p$ -value was less than 0.05. Alternative procedures based on Neyman-Pearson theory exist that are not as conservative, cf. Benjamini and Hockberg (1995).

Once a case has been identified as unusual by having, say, the largest Cook's distance, the question remains of what to do with this information. If the case can be identified as an obvious error, it should either be corrected or be deleted. If it is not obviously an error, the analysis should be re-run with that case deleted. If the main inferences do not change substantially, there is no need to delete the observation. If case deletion does alter one or more inferences in a substantial way, then the results of both analyses should be reported with emphasis on how the inferences change. The  $t(n - r - 1)$  test checks whether the observation is consistent with the model and the other data. If the observation "fails to fit" the model, it may be deleted, and emphasis placed on the analysis without that case. Since all our models are only approximations, this case somehow falls outside the range where our model gives good approximations. However, it is still important to report how the inferences would be different if that case were included in the analysis.

**EXERCISE 9.19.** For the FEV data of Exercise 9.18, construct the matrices

$$\begin{aligned} H &= X(X'X)^{-1}X', & \hat{\epsilon} &= Y - \hat{Y} = Y - X\hat{\beta}, & V &= \text{diag}(H), \\ S &= \left( \sqrt{(I - V)\text{MSE}} \right)^{-1} \hat{\epsilon}, & D &= V(I - V)^{-1}\text{diag}(S)S, \end{aligned}$$

where  $\text{diag}(S)$  is a diagonal matrix that has the elements of the vector  $S$  along the diagonal but  $\text{diag}(H)$  is a diagonal matrix that zeros out all of the entries in the square matrix  $H$  except those

Table 9.6: *FEV model selection statistics. S = Smoke, A = Age.*

Model	Predictors	LPML	DIC	BIC
(1)	S	-394.73	789.20	800.70
(2)	A	-356.53	712.60	724.08
(3)	S, A	-356.26	711.90	727.16
(4)	S, A, SA	-351.51	702.40	721.52
(5)	S, A, A <sup>2</sup> , SA, SA <sup>2</sup>	-350.71	700.30	727.18
(6)	S, A, SA, (1 - S)A <sup>2</sup>	-349.82	698.40	721.36

down the diagonal. (a) Identify the case with the largest Studentized residual and the one with the largest Cook's distance. (b) Test the hypothesis that the former case belongs to the assumed model. (c) Remove the case with the largest  $D_i$  and re-run the analysis. Report any substantial differences from the analysis done in Exercise 9.18. You may want to revisit Appendix C for a reminder of useful R commands. The website for the book contains code for calculating these matrices. (Of course, all of these features have been preprogrammed into any good frequentist regression program.)

EXERCISE 9.20. Justify that  $T_{fi}$  has the specified  $t$  distribution.

## 9.7 Model Selection

For linear regression we illustrate the use of the model selection criteria of Section 4.9, that is, the log pseudo marginal likelihood (LPML) with the corresponding pseudo Bayes factor (PBF), the deviance information criterion (DIC), and the Bayesian information criterion (BIC). Candidate models are distinguished by different covariate combinations or transformations of predictor variables. As discussed in Subsection 4.9.4, transforming the dependent variable will invalidate such comparisons between models that use different transformations unless special adjustments are made.

EXAMPLE 9.7.1. *FEV Data.* Table 9.6 gives LPMLs, DICs, and BICs for a variety of models predicting FEV from S (smoke) and A (age). Models (1) and (2) are the simple linear regression models, which in the case of model (1) is a two-group model. Model (3) is the ACOVA (parallel lines) model. Model (4) has separate straight lines for smokers and nonsmokers. Model (5) has separate quadratic trends for smokers and nonsmokers and model (6) includes a straight line trend for smokers and quadratic trend for nonsmokers. Diffuse, independent  $N(0, 1000)$  priors were used for all regression coefficients independently of a Gamma(0.001, 0.001) for the precision.

Although model (6) is supported over the other models by all three selection criteria, the values of DIC and LPML for models (4) and (5) are not very different from those for model (6). One might argue that the more parsimonious model (4) should prevail. Although a consensus was reached on model (6) by DIC, BIC, and LPML in this example, different criteria can give rise to different preferred models. The best approach to model selection combines a selection criterion with guidance from subject-matter experts.

We also computed the PBFs for all 15 comparisons among the 6 models. Model (1) without age clearly fits worse than any other model:  $PBF_{21} = 4 \times 10^{16}$ ,  $PBF_{31} = 5 \times 10^{16}$ ,  $PBF_{41} = 6 \times 10^{18}$ ,  $PBF_{51} = 1 \times 10^{19}$ ,  $PBF_{61} = 3 \times 10^{19}$ . Model (2) with age alone fits only slightly worse than the model with parallel lines for age,  $PBF_{32} = 1.31$ , but the model with age alone clearly fits worse than any of the more sophisticated models:  $PBF_{42} = 150.9$ ,  $PBF_{52} = 336.8$ ,  $PBF_{62} = 820.7$ . Since it hardly fits better than the model with age alone, the parallel lines model (3), not surprisingly, also fits much worse than any of the more sophisticated models:  $PBF_{43} = 115.3$ ,  $PBF_{53} = 257.3$ ,  $PBF_{63} = 627.0$ . The real choice is between models (4), (5), and (6) with  $PBF_{54} = 2.2$ ,  $PBF_{64} = 5.4$ ,  $PBF_{65} = 2.4$ . Model (6), which is bigger than model (4) yet smaller than model (5), is preferred.

The pseudo marginal likelihood ( $PML$ ), and its building blocks, the conditional predictive ordinates ( $CPO$ ), were discussed in Subsection 4.9.2. Recall that for a given model  $j$ , the  $PML_j$  is the product over  $i$  of the  $CPO_{ij}$  statistics and therefore the log of the  $PML_j$  is given by  $LPML_j = \sum_{i=1}^n \log(CPO_{ij})$ . The corresponding pseudo Bayes factor comparing models  $j$  and  $j'$  is  $PBF_{jj'} = \exp(LPML_j - LPML_{j'})$ .

$CPO$  and  $LPML$  statistics can be computed using WinBUGS in conjunction with R. For a given linear regression model, the  $CPO$  statistic for case  $i$  is

$$CPO_i = \left\{ E_{\beta, \tau|Y} \left( \frac{1}{p(y_i|\beta, \tau)} \right) \right\}^{-1},$$

where the values of  $CPO_i^{-1}$  can be computed directly in WinBUGS by defining WinBUGS nodes for  $p(y_i|\beta, \tau)^{-1}$ . The arithmetic needed to obtain  $CPO_i$  itself is done in R. For FEV model (4) and the BCJ prior, add the following lines to the WinBUGS code in Subsection 9.4.6:

```
for(i in 1:n){  
  CPOinv[i] <- sqrt(2*3.14159/tau)*exp(0.5*tau*pow(FEV[i]-mu[i],2))  
}
```

Note that  $CPOinv[i]$  is  $p(y_i|\beta, \tau)^{-1}$ , where the response is  $y_i = \text{FEV}_i$ . WinBUGS will output the posterior mean of each  $CPOinv[i]$ , which gives a numerical approximation to  $(CPO)_i^{-1} = E_{\beta, \tau|Y}[p(y_i|\beta, \tau)^{-1}]$ , for  $i = 1, \dots, n$ .

We've now reached the inconvenient part associated with calculating  $LPML$ . When running the model directly in WinBUGS, we need to extract the  $CPOinv[i]$ 's posterior means to compute their reciprocals, and ultimately the  $LPML$ . Moreover, we have to do this 6 times, once for each model in Table 9.6. It is more convenient to use the `bugs` function to drive WinBUGS from within R, as illustrated in Subsection 9.4.6 and Section C.6, in which case the required posterior means are automatically returned to R.

Prior to the `bugs` command in the R program in Subsection 9.4.6, add (the vector) `CPOinv` to the list of monitored parameters. Once the `bugs` command is successfully executed in R, extract the posterior mean of `CPOinv` and calculate its reciprocal to get the vector of  $CPO$  statistics. Finally, calculate  $LPML$ . R commands for both are displayed below:

```
CPO <- 1/FEV.fit$mean$CPOinv # CPO is a vector of length n  
LPML <- sum(log(CPO))
```

To reduce the computational storage burden, the R code in Section 9.4.6 was modified further by reducing the number of MCMC iterates to `n.iter=12000` with `n.burnin=2000`. Using the BCJ prior with model (4), the  $LPML$  is  $-351.02$ . From Table 9.6, the  $LPML$  when using a proper reference prior is  $-351.51$ . R and WinBUGS programs for this linear regression analysis of the FEV data are available at our website.

**EXERCISE 9.21. FEV Data Analysis Project.** Conduct a complete Bayesian linear regression analysis of the *full* FEV data available from our website that includes as candidate predictor variables Age, Smoke, Height (in inches), and Sex (1 = male). One issue to be addressed in this study is the determination of normal ranges of FEV for a given type of adolescent. For instance, what FEV is predicted for a 15-year-old male who does not smoke and is 66 inches tall? Do the following in your analysis.

- (a) An exploratory data analysis.
- (b) Discuss prior construction, predictor selection (consider interactions and higher order terms), and convergence and model diagnostics.
- (c) Present posterior inferences for regression parameters and for subpopulation means in appropriately designed tables or figures. Based on your analysis, is smoking related to FEV?
- (d) Determine normal FEV ranges for several different types of adolescents presenting the results in a table.

- (e) Discuss a sensitivity analysis.  
(f) Write-up of your entire analysis.

Dr. David Mannino, M.D. (Division of Pulmonary, Critical Care, and Sleep Medicine and Director of the Pulmonary Epidemiology Research Laboratory at the University of Kentucky) provided prior information. The values are measured in liters. The two numbers are the prior best guess of the mean FEV followed by the 99th percentile for the mean. For 18-year-old, female smokers, 70 inches tall: 4.0 and 4.8. For 16-year-old, male nonsmokers, 70 inches tall: 4.2 and 5.0. For 13-year-old, male smokers, 66 inches tall: 3.4 and 4.0. For 12-year-old, male nonsmokers, 60 inches tall: 2.7 and 3.5.

**EXERCISE 9.22.** *The Coleman Report Data.* Mosteller and Tukey (1977) and Christensen (1996) reproduced data collected from schools in the New England and Mid-Atlantic states of the USA. Consider two variables:  $y$  – the mean verbal test score for sixth graders and  $x$  – a composite measure of socioeconomic status associated with the school. The data are presented in Table 9.7. We wish to predict  $y$  based on  $x$ . Conduct a complete Bayesian regression analysis using proper reference priors. Present a scatterplot of the data with the estimated regression line and a point-wise 95% probability band. Quantify the association between  $x$  and  $y$  using an approach of your choosing. Predict the value for new schools with  $x = -16.04$ . Present posterior inferences for parameters and predictions in an appropriately designed table. Would you be surprised if higher socioeconomic status were positively associated with higher test scores?

Table 9.7: *Coleman Report data.*

School	$y$	$x$	School	$y$	$x$
1	37.01	7.20	11	23.30	-12.86
2	26.51	-11.71	12	35.20	0.92
3	36.51	12.32	13	34.90	4.77
4	40.70	14.28	14	33.10	-0.96
5	37.10	6.31	15	22.70	-16.04
6	33.90	6.16	16	39.70	10.62
7	41.80	12.70	17	31.80	2.66
8	33.40	-0.17	18	31.70	-10.99
9	41.01	9.85	19	43.10	15.03
10	37.20	-0.05	20	41.01	12.77

## 9.8 Nonlinear Regression\*

Throughout this chapter we have assumed that  $E[y|x] = x'\beta$ . We now allow a more general form for the regression function,  $E[y|x] = m(x;\beta)$  where  $m(x;\beta)$  is a known function of  $x$  and  $\beta$ . In practice,  $x$  is observed but  $\beta$  is an  $r$  vector of unknown parameters. As before, although  $x$  is often random, we condition on it and so treat it as fixed. The nonlinear regression model is

$$\begin{aligned} y_i | \beta, \tau &\stackrel{\text{ind}}{\sim} N(\theta_i, 1/\tau) \\ \theta_i &= m(x_i; \beta). \end{aligned}$$

**EXAMPLE 9.8.1.** Carlin and Gelfand (1991) reported data from a growth study by Ratkowsky (1983) on length  $y$  and age  $x$  measurements collected on 27 dugongs, a large marine mammal. The data can be found at our website, along with R and WinBUGS programs for this example. Carlin and Gelfand consider a growth curve model that is similar to

$$\begin{aligned} y_i | \beta, \tau &\stackrel{\text{ind}}{\sim} N(\theta_i, 1/\tau) \\ \theta_i &= \beta_1 - e^{\beta_2 + \beta_3 x_i}. \end{aligned}$$

For nonlinear regression, the question arises as to how to place a prior on the  $r$  vector  $\beta$ . Our method of prior elicitation is unchanged from linear regression (or even from logistic regression). We proceed by placing a prior on parameters that a scientist can think about directly. For nonlinear regression write

$$\tilde{m} \equiv m_r(\beta) \equiv m(\tilde{X}; \beta). \quad (1)$$

Here  $\tilde{m}$  is an  $r$  vector of mean values under the nonlinear model corresponding to a matrix of predictors  $\tilde{X}$ . The function  $m_r$  maps  $r$  dimensional vectors  $\beta$  into  $r$  dimensional vectors  $\tilde{m}$ . A BCJ prior is constructed by eliciting a prior

$$\tilde{m} \sim N_r(\tilde{m}_0, D(\tilde{w})) \quad (2)$$

based on information about the mean response for the  $r$  predictor vectors  $\tilde{x}_i$  that comprise the rows of  $\tilde{X}$ . The additional difficulty with nonlinear regression lies in inducing the prior on  $\beta$  from the prior on  $\tilde{m}$ .

To induce the prior on  $\beta$  from the prior on  $\tilde{m}$ , we need the inverse of the function  $m_r$ , i.e., we need to find a function  $h$  so that  $\beta = h(\tilde{m})$ . Obviously,  $m_r(h(\tilde{m})) = \tilde{m}$ . For the growth study example, standardize the ages to get  $x_i$ s and let  $\tilde{X} = (0, 1, -1)'$  correspond to the average age, one standard deviation above the average age, and one standard deviation below. Finding the inverse amounts to solving three equations

$$\begin{pmatrix} \tilde{m}_1 \\ \tilde{m}_2 \\ \tilde{m}_3 \end{pmatrix} = \begin{pmatrix} \beta_1 - e^{\beta_2} \\ \beta_1 - e^{\beta_2 + \beta_3} \\ \beta_1 - e^{\beta_2 - \beta_3} \end{pmatrix} \quad (3)$$

for  $\beta$ . If there is no obvious solution and symbolic manipulation programs like Maple or Matlab cannot solve it, we find a linear approximation to the nonlinear model that has an explicit solution. This is done using a multivariate extension of Taylor's Theorem (which is discussed in Section A.10). Given the linear approximation, we induce an approximate BCJ prior onto the solution of the linear approximation and proceed with that as our prior on the regression coefficients.

The first thing to do is to find the best guess for  $\beta$ , i.e.,  $\beta_0$  that satisfies

$$\tilde{m}_0 = m_r(\beta_0).$$

If we know  $h$ , this is trivial:  $\beta_0 = h(\tilde{m}_0)$ . More often we do not know  $h$  but the solution can be found by using a Newton iterative routine. We require the matrix of partial derivatives of  $m_r(\beta)$ , say,

$$\dot{m}_r(\beta) = \left\{ \frac{\partial m(\tilde{x}_i; \beta)}{\partial \beta_j} : i, j = 1, \dots, r \right\},$$

which is obtained by taking each row on the right-hand side of (3) and differentiating with respect to  $\beta_1$ , with respect to  $\beta_2$ , and with respect to  $\beta_3$ , to obtain a row vector of partial derivatives. We need the resulting matrix to be nonsingular. In the example we obtain

$$\dot{m}_r(\beta) = \begin{pmatrix} 1 & -e^{\beta_2} & 0 \\ 1 & -e^{\beta_2 + \beta_3} & -e^{\beta_2 + \beta_3} \\ 1 & -e^{\beta_2 - \beta_3} & e^{\beta_2 - \beta_3} \end{pmatrix}. \quad (4)$$

Write the first order Taylor expansion of the function  $m_r(\beta)$  about an arbitrary value  $\beta^*$ , which gives

$$m_r(\beta) \doteq m_r(\beta^*) + \dot{m}_r(\beta^*)(\beta - \beta^*).$$

To find the solution of  $\tilde{m}_0 = m_r(\beta_0)$ , write

$$\tilde{m}_0 \doteq m_r(\beta^*) + \dot{m}_r(\beta^*)(\beta - \beta^*),$$

solve for  $\beta$ , and repeat the process with this new  $\beta$  in place of  $\beta^*$ . If we have performed this action  $k$  times to obtain an approximate solution  $\beta^{(k)}$ , the next solution has the form

$$\beta^{(k+1)} = \beta^{(k)} + [\dot{m}_r(\beta^{(k)})]^{-1} (\tilde{m}_0 - m_r(\beta^{(k)})).$$

With luck or a nicely behaved  $m_r$  function,  $\beta^{(k+1)}$  converges to the solution  $\beta_0$ .

The iterative process begins with a starting value  $\beta^* \equiv \beta^{(0)}$ , perhaps the maximum likelihood estimate. At every stage the value  $\beta^{(k)}$  must result in a nonsingular matrix  $\dot{m}$ . In the growth study example,  $\beta = 0$  gives a singular matrix in (4) so that won't work for a starting value. To decide on when  $\beta^{(k+1)}$  is close enough to  $\beta_0$ , continue until you find a value of  $k$  for which  $\sum_{j=1}^r |\beta_j^{(k+1)} - \beta_j^{(k)}| \leq \delta$ , where  $\delta$  is a small value, perhaps 0.001.

Now that we have a best guess for  $\beta$  we will use that as the launchpad for a linear approximation to the inverse function  $h$ . Although we do not know  $h$ , we do know  $\beta_0 = h(\tilde{m}_0)$  and it turns out that we can find the derivatives of  $h$ . Begin by taking a first order Taylor expansion of  $h$  about our prior guess  $\tilde{m}_0$  to obtain

$$\beta = h(\tilde{m}) \doteq h(\tilde{m}_0) + \dot{h}(\tilde{m}_0)(\tilde{m} - \tilde{m}_0).$$

On the right-hand side  $h(\tilde{m}_0) = \beta_0$  and  $\tilde{m}_0$  are both known, as will be  $\dot{h}(\tilde{m}_0)$ , so this is a linear function that (approximately) maps  $\tilde{m}$  into  $\beta$ . Since we have placed a normal prior on  $\tilde{m}$  in (2), we obtain an induced normal prior on the linearized solution,

$$\beta \sim N_r(\beta_0, \dot{h}(\tilde{m}_0)D(\tilde{w})\dot{h}(\tilde{m}_0)').$$

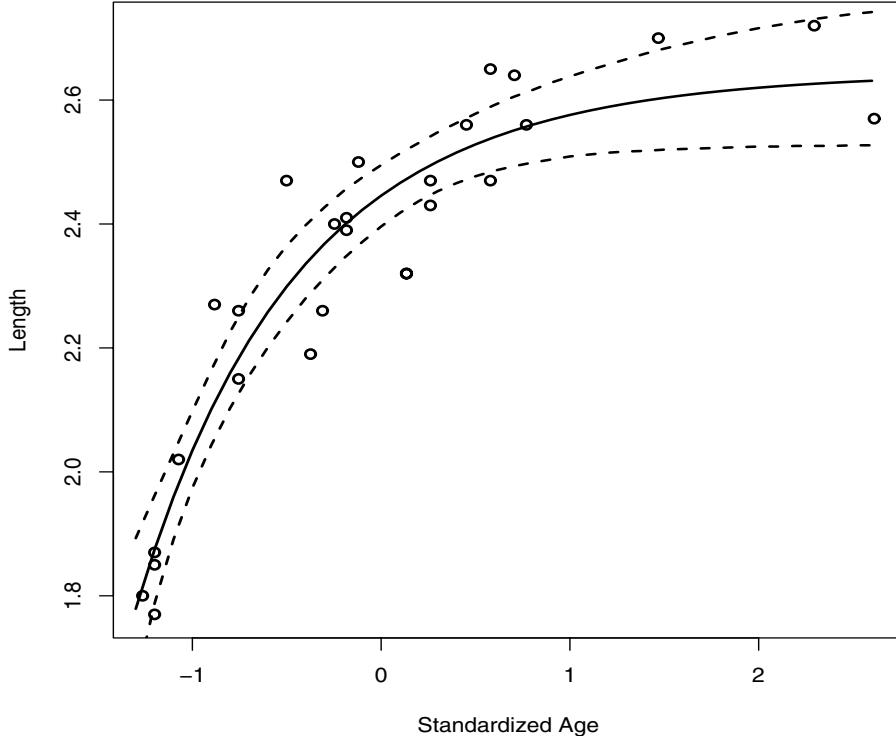


Figure 9.6: Dugong growth curve. Solid is posterior mean and dashed is 95% pointwise posterior band.

The only wrinkle is that we need to know  $\dot{h}(\tilde{m}_0)$ , the derivative of  $h$  evaluated at  $m_0$ . We assumed the existence of  $h$  but unless we can write it down, it is only implicitly defined. However, since  $m_r(h(\tilde{m})) = \tilde{m}$ , differentiating both sides we have  $\dot{m}_r(h(\tilde{m}))\dot{h}(\tilde{m}) = I_r$ , which is equivalent to  $\dot{h}(\tilde{m}) = [\dot{m}_r(h(\tilde{m}))]^{-1}$ . To obtain the term in the induced covariance matrix, evaluate (4) at the solution  $\beta_0$  and compute the inverse, thus

$$\beta \sim N_r(\beta_0, [\dot{m}_r(\beta_0)]^{-1} D(\tilde{w}) [\dot{m}_r(\beta_0)]^{-1}).$$

If a noninformative analysis is desired, we recommend determining some form of maximum range of values for some  $\tilde{m}$ s, and selecting normal distributions for the  $\tilde{m}_i$ s that cover those ranges with sufficient probability.

A final word of warning. It can be difficult to get convergence in the MCMC algorithms for nonlinear regression. You may need to run very long chains and it might help to reparameterize the problem.

**EXAMPLE 9.8.1 CONTINUED.** For the Dugong growth study, primary interest lies in the mean length as a function of age. The random variable mean length is  $\theta(x) \equiv m(x; \beta)$ . Figure 9.6 plots the expected value and the 95% PI for standardized ages  $x$  of  $-1.3$  (age 0.7),  $-1.2, \dots, 2.6$  (age 31.4). Clearly, mean length increases with a plateau around  $x = 1$  (age 19).

The main difficulties in this example involve obtaining  $\beta_0$  and the prior precision matrix. Both can be computed in R. We used  $\tilde{X} = (0, 1, -1)', \tilde{m}_0 = (2.4, 2.6, 2.0)', D(\tilde{w}) = \text{diag}(25, 25, 25)$ , and  $\delta = 0.001$ , with a Gamma(0.001, 0.001) prior for  $\tau$ . The induced prior on  $\beta$  is

$$\beta \sim N \left( \begin{bmatrix} 2.8 \\ -0.9 \\ -0.7 \end{bmatrix}, \begin{bmatrix} 825.0 & 2312.5 & 1312.5 \\ 2312.5 & 6562.5 & 3750.0 \\ 1312.5 & 3750.0 & 2187.5 \end{bmatrix} \right).$$

---

## Chapter 10

---

# Correlated Data

---

This chapter presents methods for analyzing correlated measurement data. Following an introduction in Section 1, Section 2 treats normal mixed models and Section 3 treats the more general models developed for iid multivariate normal data. Growth curve (longitudinal regression) models are introduced in Section 4. Most of our discussion is framed around longitudinal data analysis. Section 5 briefly discusses Gibbs sampling and missing or irregular data.

### 10.1 Introduction

Sampling individuals that share a common environment or taking multiple observations on a unit typically generates dependent observations. In particular, cross-sectional studies that involve multi-stage sampling result in dependent outcomes due to *clustering* of individuals. For example, suppose we sample hospitals, then sample patients within each hospital and measure patient satisfaction. Given their hospital, patients may look independent. But patients in a hospital share the same staff and hospital policies, so, relative to other hospitals, their similar conditions result in dependent outcomes. We usually assume independence of individuals given (conditional on) their cluster (hospital) and independence among clusters. Unconditionally, individuals within a cluster are correlated.

Longitudinal data (sometimes called panel data) arise when *repeated measurements* are taken on each unit at various times. For example, in a sample of hospitals, we could obtain overall measures of patient satisfaction but then collect those measurements at several times. Longitudinal studies are often conducted to assess time trends, and the effect of covariates on time trends. For instance, a trial designed to compare different toenail fungus treatments might involve a baseline (pre-treatment) measurement on each person prior to treatment assignment, followed by several post-assignment measurements. We may expect different trends in fungus eradication over time depending on the treatment. Units are assumed to be independent but the repeated measurements are typically dependent.

A balanced longitudinal sampling design collects the same number of observations at the same time points for each subject. A convenient special case has the observations taken at equally spaced time points. Imbalance complicates modeling; it occurs when sampling times vary among subjects.

A common phenomenon in longitudinal, spatial, and spatiotemporal data is that the correlation between observations decays with increased differences in time or distance. Spatial data collected longitudinally are called spatiotemporal data. We do not discuss spatial data but recommend the excellent treatment by Banerjee, Carlin, and Gelfand (2004).

**EXAMPLE 10.1.1. *Reference Values for IL-1 $\beta$ .*** When we were growing up, our mothers always told us that if we did not keep our interleukin-1 $\beta$  levels low, our teeth would fall out. In support of this old wives' tale, Thomas et al. (2009) reported data from a longitudinal study on the variability of several potential biomarkers as measured on healthy adults. These included the concentration in picograms per milliliter of interleukin-1 $\beta$  (IL-1 $\beta$ ), a pro-inflammatory cytokine protein that, among other clinical uses, has been suggested as a measure of periodontal health. Knowledge of the natural variability in healthy subjects allows us to detect abnormally high values and screen for adverse health conditions.

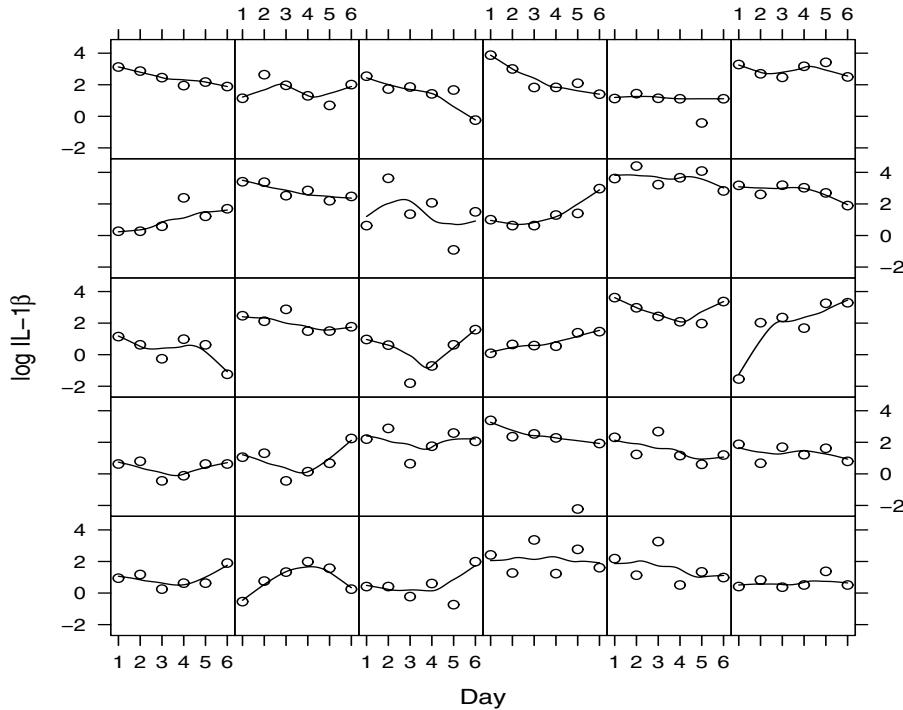


Figure 10.1: Trellis plot containing each individual's  $\text{IL-1}\beta$  data vector on the log scale.

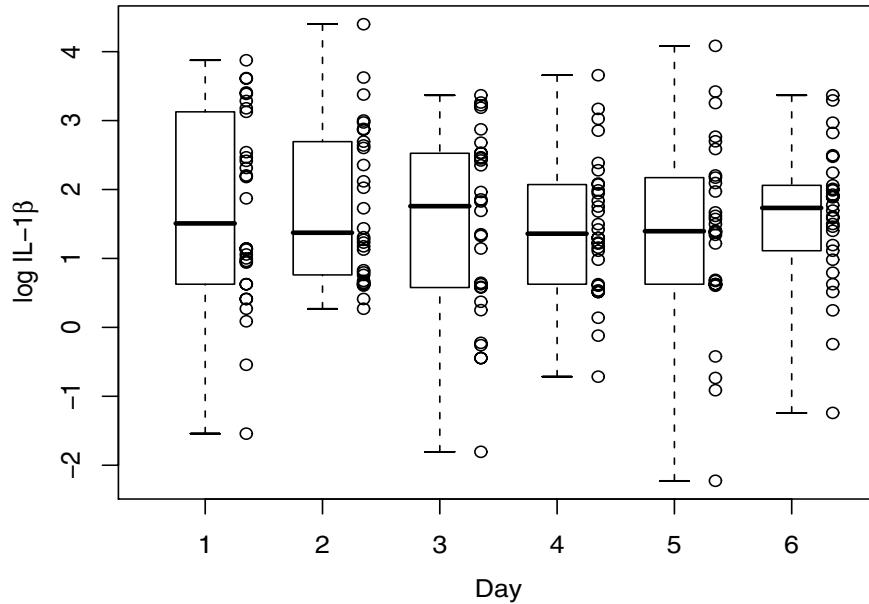
$\text{IL-1}\beta$  concentrations were measured from saliva samples collected on 6 consecutive days for 30 subjects. Figure 10.1 presents the log transformed data vectors  $y_i = (y_{i1}, y_{i2}, y_{i3}, y_{i4}, y_{i5}, y_{i6})$  in a trellis plot with LOWESS trend lines. Note the clear outlier (interleukin interloper?) on day 5 for subject 22 (row 4, column 4). Boxplots and line plots of the data appear reasonably symmetric with constant mean and variance over time, cf. Figure 10.2. See also the spaghetti plot in Figure 10.3.

The sample correlation matrix for the data is

$$\hat{R} = \begin{pmatrix} 1.00 & 0.62 & 0.57 & 0.51 & 0.26 & 0.18 \\ 0.62 & 1.00 & 0.58 & 0.71 & 0.34 & 0.46 \\ 0.57 & 0.58 & 1.00 & 0.68 & 0.42 & 0.28 \\ 0.51 & 0.71 & 0.68 & 1.00 & 0.44 & 0.37 \\ 0.26 & 0.34 & 0.42 & 0.44 & 1.00 & 0.30 \\ 0.18 & 0.46 & 0.28 & 0.37 & 0.30 & 1.00 \end{pmatrix}.$$

An independence model seems clearly inappropriate.

Dependence among data exists in many forms. Measurements of different variables taken on the same individual at the same time, like systolic blood pressure, diastolic blood pressure, and cholesterol, would generally be dependent. There is little reason to expect any structure among their covariances. Dependence also exists when the same variable is measured on the same individual repeatedly in time. For example, if we measure an individual's cholesterol over time, we expect those measurements to be dependent. In particular, we expect them to be more highly correlated the closer the proximity in time. Dependence also occurs based on group membership. Health outcome measurements taken on animals from the same farm are often dependent because the animals share the same living circumstances, e.g., management practices, food, etc. In this case, a reasonable model for each outcome has the same correlation among all animals from a given farm. This last

Figure 10.2: Boxplots of  $\log(IL-1\beta)$ .

case is well represented by the mixed models of the next section. All the cases are amenable to the methods of Section 2. Determining an appropriate model depends on the structure of the data.

## 10.2 Mixed Models

Our usual sampling model for regression is

$$y_k = x'_k \beta + \varepsilon_k, \quad \varepsilon_k | \sigma \stackrel{iid}{\sim} N(0, \sigma^2), \quad k = 1, \dots, n.$$

Here we have sampled  $n$  individuals and on each have observed a single dependent variable  $y_k$  along with a vector containing predictor information  $x_k$ . Using a slight change of notation to help introduce mixed models, write the predictor variables as two vectors  $x_k$  and  $z_k$  of dimensions  $r$  and  $q$ , respectively. If we similarly write the regression coefficients as two vectors  $\beta$  and  $b$ , we can write a regression model as

$$y_k = [x'_k, z'_k] \begin{pmatrix} \beta \\ b \end{pmatrix} + \varepsilon_k = x'_k \beta + z'_k b + \varepsilon_k.$$

The point is simply that we can split the linear model into two parts corresponding to different regressors and their corresponding coefficients.

To get a mixed model, we make  $b$  random. Generalizing the fixed effects linear model of Section 9.1, write the general mixed model as

$$Y = X\beta + Zb + e.$$

Our discussion focuses on grouped data, just as it did in Section 8.5 on binomial regression mixed models, but we introduce some more general models than were considered there.

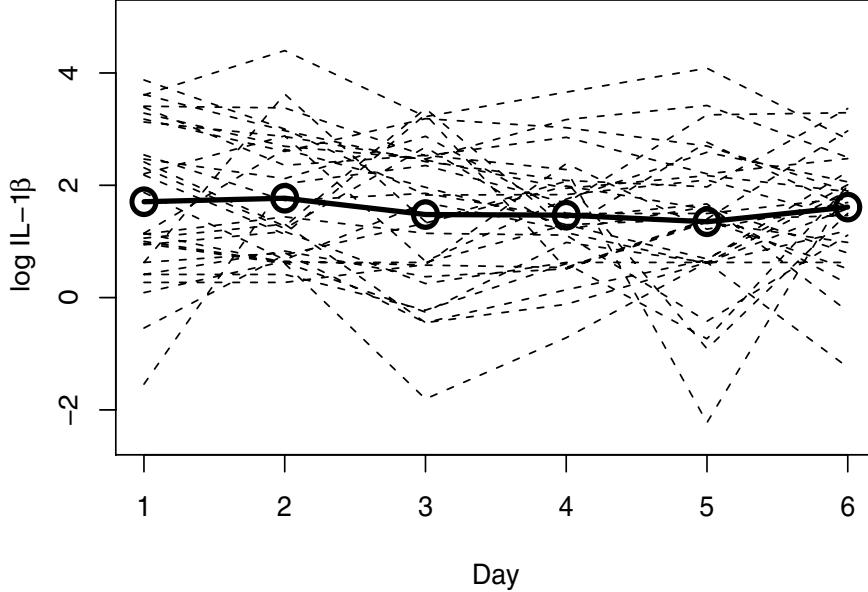


Figure 10.3: Spaghetti plot of  $\log(IL-1\beta)$ , with sample means.

Suppose we have groups  $i = 1, \dots, a$  with  $n_i$  observations on group  $i$ . These might be repeated observations on an individual. We focus on models with a separate random effects vector  $b_i$  for each group,

$$y_{ij} = [x'_{ij}, z'_{ij}] \begin{pmatrix} \beta \\ b_i \end{pmatrix} + \varepsilon_{ij} = x'_{ij}\beta + z'_{ij}b_i + \varepsilon_{ij}, \quad j = 1, \dots, n_i. \quad (1)$$

Here  $y_{ij}$  is the  $j$ th response from group  $i$ ; perhaps the response of individual  $i$  at time  $t_{ij}$  or the response of the  $j$ th individual in cluster  $i$ . Henceforth, we refer to unit  $i$ , which may be an individual or a group.

With  $\beta$  an  $r$  vector and each  $b_i$  a  $q$  vector, collect all responses from unit  $i$  into a vector and write

$$y_i \equiv \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{in_i} \end{pmatrix} = \begin{pmatrix} x'_{i1} \\ x'_{i2} \\ \vdots \\ x'_{in_i} \end{pmatrix} \beta + \begin{pmatrix} z'_{i1} \\ z'_{i2} \\ \vdots \\ z'_{in_i} \end{pmatrix} b_i + \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \vdots \\ \varepsilon_{in_i} \end{pmatrix} \equiv X_i \beta + Z_i b_i + \varepsilon_i.$$

Here we allow the predictor information to differ both within and between units.

In the sampling model for standard regression, all coefficients are fixed unknown parameters. In a mixed model,  $\beta$  is fixed but the  $b_i$ 's are random. When  $x'_{ij}\beta$  includes an intercept, a standard assumption is

$$b_i | \xi \stackrel{iid}{\sim} N_q(0, \Sigma_b(\xi)), \quad \perp \!\!\! \perp \quad \varepsilon_i | \sigma \stackrel{ind}{\sim} N_{n_i}(0, \sigma^2 I_{n_i}),$$

with  $\Sigma_b(\xi)$  a known positive definite matrix given the value of the parameter vector  $\xi$ . We can also think of this as a mixture (hierarchical) model by writing

$$y_i | \beta, b_i, \sigma^2 \stackrel{ind}{\sim} N_{n_i}(X_i \beta + Z_i b_i, \sigma^2 I_{n_i}), \quad b_i | \xi \stackrel{iid}{\sim} N_q(0, \Sigma_b(\xi)).$$

To define the matrix model for all the data, write

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_a \end{pmatrix} \quad X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_a \end{pmatrix} \quad Z = \begin{pmatrix} Z_1 & 0 & \cdots & 0 \\ 0 & Z_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Z_a \end{pmatrix} \quad b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_a \end{pmatrix}.$$

All the matrices have  $n$  rows where  $n = \sum_{i=1}^a n_i$  is the total number of observations. The matrix  $Z$  is a *block diagonal* matrix. The block diagonal takes the matrices  $Z_i$  and places them down the diagonal to create an  $n \times aq$  matrix with zeros everywhere off the block diagonal terms. Our general mixed model for grouped data is

$$Y = X\beta + Zb + e, \quad b_i | \xi \stackrel{iid}{\sim} N_q(0, \Sigma_b(\xi)) \quad \perp \quad e | \sigma^2 \sim N_n(0, \sigma^2 I_n). \quad (2)$$

In a common special case, fixed effects predictors do not vary within units so  $x_{ij} = x_i$  for all  $j$ . This causes the rank of  $X_i$  to be 1 but  $X$  can remain a full rank matrix. Standard operations on mean vectors and covariance matrices establish a marginal sampling distribution

$$Y | \beta, \sigma, \xi \sim N_n(X\beta, Z[\text{block diagonal}(\Sigma_b(\xi))]Z' + \sigma^2 I_n). \quad (3)$$

Again, we can also think of this as a mixture model by writing

$$Y | \beta, b, \sigma, \xi \sim N_n(X\beta + Zb, \sigma^2 I_n), \quad b_i | \xi \stackrel{iid}{\sim} N_q(0, \Sigma_b(\xi)).$$

Our discussion is restricted to this model for independent data units. Great generalizations are possible. For example, there is little reason why  $q$ , i.e., the length of  $z_{ij}$ , should not depend on the unit  $i$ . Also, the error variance can vary across units (e.g., different variance for different treatment groups). Moreover, the general mixed model allows units to be correlated as well as responses within units. This involves a non-block-diagonal design matrix  $Z$  and different assumptions about  $b$  than used here. In fact, mixed models can even be used to analyze spatial data.

### 10.2.1 Random Intercept Model

Consider the sampling model

$$y_{ij} = x'_{ij}\beta + b_i + \varepsilon_{ij}; \quad b_i | \tau_b \stackrel{iid}{\sim} N(0, 1/\tau_b) \quad \perp \quad \varepsilon_{ij} | \tau \stackrel{iid}{\sim} N(0, 1/\tau) \quad (4)$$

$i = 1, \dots, a$  and  $j = 1, \dots, n_i$ . Here,  $q = 1 = z_{ij}$  and  $x'_{ij}\beta$  includes an intercept. For longitudinal data,  $y_{ij}$  is the response and  $x_{ij}$  is the covariate vector recorded at time  $j$  on unit  $i$ . Often the covariate vector includes the time of the observation  $t_{ij}$  or functions of it. The random intercept  $b_i$  represents a latent unit-specific effect that quantifies the degree to which unit  $i$  responds above ( $b_i > 0$ ) or below ( $b_i < 0$ ) the subpopulation mean  $x'_{ij}\beta$ . Special cases include the one-way random effects model, more general ANOVA models with one random effect, and straight line regression with random intercepts, all of which will be discussed. Section 8.5 focused on the binomial regression version of this model.

In model (4) with  $\sigma_b^2 = 1/\tau_b$  and  $\sigma^2 = 1/\tau$ , the covariance structure is

$$\text{Cov}(y_{ij}, y_{i'j'}) = \begin{cases} \sigma^2 + \sigma_b^2 & \text{if } (i, j) = (i', j') \\ \sigma_b^2 & \text{if } i = i' \text{ and } j \neq j' \\ 0 & \text{if } i \neq i' \end{cases}.$$

This is known as a *compound symmetry* or *intraclass correlation* structure. The variances of all observations are the same. Responses for different units  $i$  (e.g., different hospitals in multistage

sampling or different people in longitudinal studies) are uncorrelated. The correlation between different measurements taken within the same unit is positive, namely

$$\rho \equiv \text{Corr}(y_{ij}, y_{ij'}) = \frac{\sigma_b^2}{\sigma^2 + \sigma_b^2}.$$

To analyze model (4), take an independence prior  $p(\beta, \tau_b, \tau) = p(\beta)p(\tau_b)p(\tau)$  with  $\beta \sim N(\beta_0, \Sigma_0)$ . The prior on the precision  $\tau$  is often taken as a Gamma distribution, but it is also common to place proper uniform priors on  $\sigma$  and  $\sigma_b$ . In the latter case, the ranges must be suitably large. Although a Gamma( $\epsilon, \epsilon$ ) prior with small  $\epsilon$  is common for  $\tau$ , Gelman (2006) argued against the routine use of this prior for variance components in random effects distributions. We often use proper uniform priors on standard deviations of random effects. See Subsection 9.4.2 for further discussion of informative priors on precisions. Informative priors can be placed on the fixed regression coefficients  $\beta$  exactly as was done in Subsection 9.4.1. Simply fix  $b_i = 0$  when making the elicitation and think of “typical” units with the specified  $\tilde{x}_{ij}$ s.

With a small number of units  $a$ , it may be difficult to obtain a precise estimate of  $\sigma_b$ . Trying to estimate the standard deviation of observables based on a small sample size is already difficult. Attempting to estimate the standard deviation of unobserved  $b_i$ s based on a small number of units will be at least as hard. This difficulty may present itself indirectly in the posterior analysis through difficulty in achieving convergence of the Markov chain or some other form of instability. In our experience, using some real prior information on  $\sigma_b$  or the corresponding precision  $\tau_b$  mitigates this problem.

Our first special case is the one-way random effects ANOVA model. It specifies a constant mean across units but a random effect  $b_i$  for each unit. It has two sources of variability about the mean, the between-unit variability  $\sigma_b^2$  and the variability of data taken within units  $\sigma^2$ . The form of (4) in this case is

$$y_{ij} = \mu + b_i + \varepsilon_{ij}. \quad (5)$$

This mean structure is appropriate for longitudinal data that fluctuate about a unit-specific mean over time, and where all of the data fluctuate about an overall average,  $\mu$ . This is a possible model for the IL-1 $\beta$  biomarker study of Example 10.1.1.

**EXERCISE 10.1.** (a) What are  $X_i$  and  $Z_i$  for model (5)? (b) What is the marginal distribution of the full data vector  $Y$  for model (5)? Simplify as best you can.

**EXERCISE 10.2.** Argue that the hierarchical model

$$y_{ij} | \mu_i, \tau \stackrel{\text{ind}}{\sim} N(\mu_i, 1/\tau) \quad \perp\!\!\!\perp \quad \mu_i | \mu, \tau_b \stackrel{\text{iid}}{\sim} N(\mu, 1/\tau_b) \quad (6)$$

is equivalent to model (5). By equivalent, we mean that it produces the same sampling distribution given the parameters  $\mu$ ,  $\tau_b$ , and  $\tau$ .

Despite the equivalence of model (5) and its hierarchical version model (6), the Markov chains obtained from Gibbs sampling will be different for the two versions. The chain for the hierarchical version may be more efficient due to less correlation among its components. Estimates of all fixed parameters will be identical up to MC error.

In analyzing the one-way random effects model, interest lies primarily in  $\mu$ , the average of the distribution of population means; in  $\tau_b$ , the precision of the random effects; and, as always, in the predictive distribution of a new observation, say  $y_f$ . If  $\tau_b$  is large, there will be little difference in the behavior of the various units. Rather than examining  $\tau_b$  directly, we find it more illuminating to look at the distribution of a new conditional mean for observations,  $\mu_f$ , sampled independently from the distribution of conditional means, namely

$$\mu_f \equiv \mu + b_f \sim N(\mu, 1/\tau_b).$$

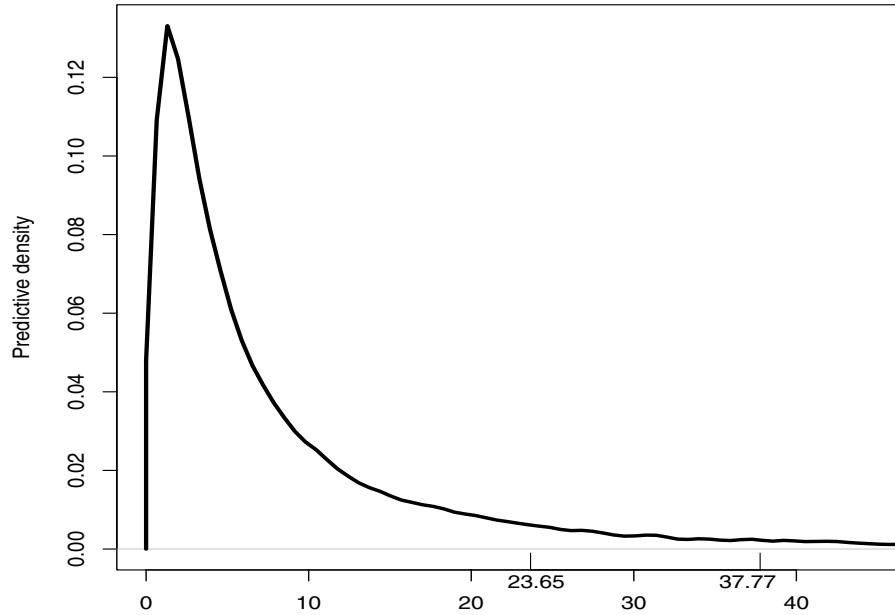


Figure 10.4: *Predictive density for future IL-1 $\beta$  score.*

We could look at percentiles, perhaps  $\gamma_{0.1} \equiv \mu - 1.28\sqrt{1/\tau_b}$  and  $\gamma_{0.9} \equiv \mu + 1.28\sqrt{1/\tau_b}$ , the 10th and 90th percentiles of the distribution. Sampling  $\mu_f$  conditional on everything else results in an estimate of the  $N(\mu, 1/\tau_b)$  distribution, which is the modeled distribution of the means. If the distribution is highly concentrated, then we know that  $\tau_b$  is very large.

Although the  $\mu_i$ s are unobserved samples from a  $N(\mu, 1/\tau_b)$  distribution, their individual predictive distributions will differ from  $\mu_f$  and each other. For  $\mu_f$  the data only give information about  $\mu$  and  $\tau_b$ , but for any  $\mu_i$  the data  $y_{ij}$ ,  $j = 1, \dots, n_i$  bear specifically on the value of  $\mu_i$ . In particular, the sample means  $\bar{y}_i$  can be used as a diagnostic tool to check whether the distribution of random effects is actually normal. Similarly, individual  $\mu_i$ s that have atypical posterior distributions, notably atypical posterior means or medians, are of interest.

**EXAMPLE 10.2.1. Reference Values for IL-1 $\beta$ .** We fit model (6) to the log transformed IL-1 $\beta$  data using a proper diffuse prior as indicated in Exercise 10.3.

As with the Diasorin data in Examples 5.2.2 and 9.5.1, we could present the IL-1 $\beta$  analysis on either the log scale or the original scale. There are advantages to each. On the log scale, distributions tend to be more symmetric, which makes interpretations easier. The original scale has the advantage of being the original scale, so presumably is more intuitive to people familiar with IL-1 $\beta$ . However, on the original scale the distributions are highly skewed, making results harder to interpret. It is worthwhile to examine both scales (although even the authors differ on their relative merits).

The data were collected to characterize “normal” levels of IL-1 $\beta$ . Clearly, the most relevant inference is the predictive distribution for a new score, on either the log or original scale. Figure 10.4 gives the highly skewed predictive density on the original scale. There is a 10% chance of exceeding 23.65 and a 5% chance of exceeding 37.77. Values above these may be cause for concern.

Table 10.1 gives a variety of posterior results, much of it on both scales. The standard deviations  $\sigma$  and  $\sigma_b$  are of the same magnitude. There is roughly the same amount of variability among people

Table 10.1: Posterior results for IL-1 $\beta$  data.

Parameter	median	95% PI		Parameter	median	95% PI	
$\sigma$	0.91	0.82	1.03				
$\sigma_b$	0.82	0.60	1.15				
$\rho$	0.45	0.28	0.62				
$\mu$	1.56	1.23	1.89	$e^\mu$	4.77	3.42	6.65
$y_f$	1.55	-0.92	4.03	$e^{y_f}$	4.71	0.39	56.54
$\mu_f$	1.55	-0.16	3.28	$e^{\mu_f}$	4.72	0.85	26.60
$\gamma_{0.1}$	0.51	-0.04	0.92	$e^{\gamma_{0.1}}$	1.67	0.96	2.51
$\gamma_{0.9}$	2.61	2.21	3.17	$e^{\gamma_{0.9}}$	13.64	9.09	23.86

as there is in the observations made on one individual. There is roughly twice as much variability when examining two observations made on different people than when they are made on the same person.

On the original scale, we are 95% certain that the population median IL-1 $\beta$  score,  $e^\mu$ , is between 3.42 and 6.65; the point estimate is 4.77. We are also 95% certain that the median for a new person,  $e^{\mu_f}$ , is between 0.85 and 26.60 with a point estimate of 4.72, and that an observation taken on a new person,  $e^{y_f}$ , is 95% certain to be between 0.39 and 56.54.

From Table 10.1, our best *a posteriori* estimate is that 90% of all new people will have a score above 1.67; our uncertainty about that estimate is captured by the 95% PI (0.96, 2.51). Similarly, we infer that 90% of all new people will have a score below 13.64 with corresponding PI (9.09, 23.86). This gives some idea of the variability in the score levels.

EXERCISE 10.3. Modify the following WinBUGS code to reproduce the results of Example 10.2.1. See `mu`, `sigma`, and `sigmab` in the code for the prior specifications. The variable `ID[]` identifies the 30 distinct individuals in the data. Each individual gets their own random effect.

```

model{
  for(k in 1:180){
    logy[k] <- log(y[k])
    logy[k] ~ dnorm(m[k],tau)
    m[k] <- mu + b[ID[k]]
  }
  for(i in 1:30){b[i] ~ dnorm(0,taub)}
  sigma ~ dunif(0,100)
  tau <- 1/(sigma*sigma)
  sigmab ~ dunif(0,100)
  taub <- 1/(sigmab*sigmab)
  mu ~ dnorm(0,0.001)
  rho <- tau/(taub+tau)
  bf ~ dnorm(0,taub)
  ef ~ dnorm(0,tau)
  muf <- mu + bf
  logyf <- muf + ef
  # Predictive density
  yf <- exp(logyf)
  # 90th and 10th percentiles for muf
  U <- mu + 1.28*sqrt(1/taub)
  L <- mu - 1.28*sqrt(1/taub)
}
ID[] y[]

```

```

1      22.8
1      17.6
remaining 178 data lines go here
END

```

**EXERCISE 10.4.** *Visualizing the One-way Random Effects Model.* Consider the one-way random effects model (5) for balanced longitudinal data taken at 5 time points on 5 units. Draw pictures of how the data might appear (i.e.,  $y$  versus  $t$  for all units) if  $\tau_b$  is much larger than  $\tau$ , and vice versa. Similarly, draw pictures when both variance components are expected to be small and again when both are large.

**EXERCISE 10.5.** For the one-way random effects model (5), show that the correlation between two observations on the same individual is  $\sigma_b^2/(\sigma^2 + \sigma_b^2)$ .

**EXERCISE 10.6.** Derive the full conditionals for  $(\mu, \tau, \tau_b, b)$  under model (5) using the partially proper independence prior

$$p(\mu)p(\tau_b)p(\tau) \propto \tau_b^{a_b-1} e^{-\tau_b b_b} (1/\tau).$$

Hint: The full conditionals for  $\tau$  and  $\tau_b$  are Gamma. The full conditional for  $\mu$  is normal. It will help if you write  $[y_{ij} - \mu - b_i] = [y_{ij} - b_i] - \mu \equiv [y_{ij}^* - \mu]$ . Then show that  $\sum_{i=1}^a \sum_{j=1}^{n_i} (y_{ij}^* - \mu)^2 = N(\bar{y}_{..} - \mu)^2 + \sum_{i=1}^a \sum_{j=1}^{n_i} (y_{ij}^* - \bar{y}_{..})^2$ , where  $\bar{y}_{..} = \sum_{i=1}^a \sum_{j=1}^{n_i} y_{ij}^*/n$ . Finally, observe that the full conditional for  $b$  involves independent normal contributions and find the conditional for  $b_i | Y, \tau, \tau_b, \mu$ . The complete the square formula of Exercise 5.22 should help. Carefully justify all steps.

More general mixed effects ANOVA models allow means to vary according to one or more categorical fixed effects. For example, a randomized block design might involve giving different drugs  $j$  to individuals  $i$ . A model that has mean response  $\mu_j$  under drug  $j$  and random individual effects  $b_i$  is

$$y_{ij} = \mu_j + b_i + \varepsilon_{ij}. \quad (7)$$

This model allows for larger or smaller overall responses under any given drug  $j$  depending on an individual's particular latent susceptibility  $b_i$ . An individual with a large positive  $b_i$  tends to have larger responses regardless of which drug they take. A *typical* individual is considered to have  $b_i = 0$ .

**EXAMPLE 10.2.2.** Table 10.2 contains data on the concentration of plasma epinephrine in 10 dogs under 3 anesthetics: Isofluorane, Halothane, and Cyclopropane. Epinephrine is essentially adrenaline and, during anesthesia, it is not desirable to have high levels of adrenaline or epinephrine. The data come from a study by Perry et al. (1974) as reported in Rice (1995).

Table 10.2: *Dog data.*

Dog	Dog	Dog	Dog	Dog	Dog	Dog	Dog	Dog	Dog	
1	2	3	4	5	6	7	8	9	10	
I	0.28	0.51	1.00	0.39	0.29	0.36	0.32	0.69	0.17	0.33
H	0.30	0.39	0.63	0.68	0.38	0.21	0.88	0.39	0.51	0.32
C	1.07	1.35	0.69	0.28	1.24	1.53	0.49	0.56	1.02	0.30

Dogs constitute a blocking factor. The objective this afternoon is to compare the 3 anesthetics. (How do we know you are reading this in the afternoon? When else but on a “dog data afternoon?”)

We expect repeated observations on the same dog to be correlated. Each dog has their own physiological and biological characteristics, so their epinephrine levels should tend to be larger or smaller across all three anesthetics.

With  $i = 1, \dots, 10$ ,  $j = 1, 2, 3$ , an appropriate sampling model for concentration  $y_{ij}$  is:

$$y_{ij} = \beta_1 I_j + \beta_2 H_j + \beta_3 C_j + b_i + \varepsilon_{ij}; \quad b_i | \tau_b \stackrel{iid}{\sim} N(0, 1/\tau_b) \perp\!\!\!\perp \varepsilon_{ij} | \tau \stackrel{iid}{\sim} N(0, 1/\tau). \quad (8)$$

Each predictor is an indicator variable for the corresponding anesthetic.

To see what happens if the correlation within dogs is ignored, we first consider the following *fixed* two-factor ANOVA model using methods from Chapter 9:

$$y_{ij} = \beta_1 I_j + \beta_2 H_j + \beta_3 C_j + \beta_4 D1_i + \dots + \beta_{13} D10_i + \varepsilon_{ij}.$$

All the predictors are indicator variables; for instance,  $D4$  equals 1 for dog 4 and zero otherwise. The mean response is modeled by the additive effect of anesthetic  $j$  and dog  $i$ . The model is over-parametrized and so it is typical to place a constraint on one of the sets of parameters. We set  $\beta_3 = 0$  to resolve this issue. Inferences from this model are compared to those from a mixed effects analysis based on model (8) in Exercise 10.7.

The fixed effects analysis used independent  $N(0, 100)$  priors for the regression coefficients and a Gamma(0.01, 0.01) prior for  $\tau$ . Posterior medians and 95% intervals for the difference in mean concentration are: for  $H - I$ , 0.03 (-0.30, 0.36), for  $C - I$ , 0.42 (0.09, 0.75), and for  $C - H$ , 0.38 (0.05, 0.71). These estimates reveal that the concentration of plasma epinephrine tends to be the highest under cyclopropane, but there may be no statistical difference between isoflurane and halothane.

The question now is, how do these inferences change when random effects for dogs replace fixed effects? For randomized blocks designs, the blocks (the 10 dogs in this example) are often viewed as a random sample from a population (of dogs in this example).

**EXERCISE 10.7. Randomized Block Design.** Using the same prior as in Example 10.2.2 and incorporating a proper but diffuse uniform prior for  $\sigma_b = 1/\sqrt{\tau_b}$ , analyze the data from Table 10.2 using a random effect for each dog as in model (8). Estimate the treatment means and their differences. Obtain estimates of the random effects. Interpret the estimated random effects to determine good doggies. Analyze the data and include WinBUGS code and output in your report. Compare and contrast the analysis and interpretations with the fixed effects analysis in Example 10.2.2.

A randomized complete block design has every treatment observed within every block, so in (7)  $i = 1, \dots, a$ ,  $j = 1, \dots, T$ , where  $T$  is the number of treatments. In this case, the analysis of treatment contrasts like  $\mu_1 - \mu_3$  hardly depends on whether the  $b_i$ s are treated as fixed or random in the sampling model. All the information on  $\mu_1 - \mu_3$  from the data is contained in

$$y_{i1} - y_{i3} = \mu_1 - \mu_3 + (\varepsilon_{i1} - \varepsilon_{i3}); \quad i = 1, \dots, a,$$

which does not depend on  $b_i$ . Except for differences in prior specifications, the Bayesian analysis of such a parameter should not depend much on whether the block effects are fixed or random. Even proper reference priors are similar with, say,  $b_i \stackrel{iid}{\sim} N(0, 1000)$  for fixed effects and for random effects  $b_i | \tau_b \stackrel{iid}{\sim} N(0, 1/\tau_b)$  with a prior  $\sigma_b = 1/\sqrt{\tau_b} \sim U(0, B)$  for sufficiently large  $B$ .

For incomplete data, that is, incomplete block designs such as balanced and partially balanced incomplete block designs (see Christensen, 2002, Sections 9.4 and 12.11 or Christensen 1996, Chapter 16), the random effect and fixed effect models give different analyses. There is information on  $\mu_1$  and  $\mu_3$  to be obtained from blocks that contain one but not both of the treatments and that information is affected by the model for the block effects.

Random block effects are typically viewed as the more appropriate model, but normally distributed block effects are not necessarily appropriate. Fixed block effects provide an approach to

dealing with nonnormality. A random effects analysis similar to the one presented later in Example 15.1.5 provides an approach to dealing with nonnormal random block effects.

**EXERCISE 10.8.** *Fuel Flow Rate Data.* Analyze the data in Table 10.3 from Scheffé (1959, p. 289) involving measurements of fuel flow rates as determined by five operators for three different nozzle types. Each operator recorded three measurements for each nozzle type. Operators are to be regarded as a random sample from a larger population of operators. The main interest here is in examining the possible differences in nozzle types. Before fitting a model to the data, obtain relevant graphical and numerical descriptive statistics to assess the research goal. The data were originally from Hicks (1956). There are some negative “rates,” so presumably zero corresponds to some sort of average or expected rate, and the numbers in the data indicate whether the operator had an above or below average rate on each try.

Table 10.3: *Fuel flow rate data.*

Nozzle	1	2	3	4	5
1	6, 6, -15	26, 12, 5	11, 4, 4	21, 14, 7	25, 18, 25
2	13, 6, 13	4, 4, 11	17, 10, 17	-5, 2, -5	15, 8, 1
3	10, 10, -11	-35, 0, -14	11, -10, -17	12, -2, -16	-4, 10, 24

A third special case of model (4) is the simple linear regression model for longitudinal data with a random intercept for each unit but fixed constant slope effects in time:

$$y_{ij} = \beta_1 + b_i + \beta_2 t_{ij} + \varepsilon_{ij}. \quad (9)$$

Other functions of time like polynomials could be used. In this random effects ACOVA model, the random intercept ( $\beta_1 + b_i$ ) is different for each unit, but the fixed rate of increase or decrease over time ( $\beta_2$ ) is the same.

**EXERCISE 10.9.** *Special Cases of the Random Intercept Model.* (a) Give a scenario (perhaps from your field of study) where each of models (5), (7), and (9) could be used for data analysis. Clearly define the response variable and the predictor variables, explaining why correlation would be expected among the responses. (b) Write a mixed effects ANOVA model of the form (4) for longitudinal data on  $n$  individuals with two fixed categorical predictors, Sex and Age (old, middle age, young), their interaction, and a random effect for each individual.

Our final example involves *meta-analysis* using a modification of the one-way random effects model. Meta-analysis refers to pooling information across multiple studies each designed to address the same scientific question. The goal is often to estimate a single effect measure common to all studies. The data frequently take the form of a summary statistic obtained from each study, e.g., a mean, odds ratio, or relative risk, and the associated frequentist standard error (estimate of the standard deviation).

**EXAMPLE 10.2.3.** *Meta-Analysis of ICU Treatments.* Patients in intensive care units (ICUs) may receive treatment to decontaminate their digestive tract. This aids in preventing infections that can lead to death. In 14 clinical trials, patients were randomized to either a dual treatment regimen involving the use of both topical and systemic antibiotics, or they were assigned to a control group. The goal is to evaluate the dual treatment protocol on reduction of mortality. We use the estimated log odds ratios from these 14 trials to estimate the population median odds ratio of death under treatment relative to death under the control. Odds ratios were discussed in Section 5.1.4. The data are included in the code for Exercise 10.10.

Our data  $y_i$  are observed log odds ratios from various studies. We assume that

$$y_i \stackrel{iid}{\sim} N(\eta_i, 1/\tau_i)$$

with  $\tau_i$  known. Thus,  $\eta_i$  is the true log odds ratio associated with study  $i$ . Knowing  $\tau_i$  is, of course, an approximation with  $\tau_i$  actually obtained from the estimated standard deviation reported for  $y_i$  in the  $i$ th study. We assume a random effects model,

$$\eta_i \stackrel{iid}{\sim} N(\mu, 1/\tau_b).$$

This means that the true log odds ratios for studies vary about an overall log odds ratio mean  $\mu$ . Finally, we specify a proper reference prior

$$\mu \sim N(0, 1000), \quad \perp \!\!\! \perp \quad \sigma_b \sim \text{Uniform}(0, 100).$$

Ultimately, we want to make inferences about  $e^\mu$ , the median odds ratio across studies. Half of the study true odds ratios will be less than  $e^\mu$  and half above.

The median odds ratio  $e^\mu$  was estimated to be 0.82 (posterior median) with 95% posterior interval for the median of (0.66, 1.01). Thus, we estimate that in half of all studies that have been or might be performed, the odds of dying with the treatment are only 82% of the odds with the control. With 95% probability, that number may be as low as 2/3s or as high as 1 (i.e., no difference). More precisely, the posterior probability that  $e^\mu$  is less than 1 is 0.97, so we are 97% sure that the majority of all such studies show improved mortality prospects. This constitutes fairly strong evidence of improved odds of survival under the dual treatment. The estimated 10th and 90th percentiles of the induced distribution of odds ratios are 0.67 and 1.01, respectively. This means that we would estimate that about 10% of odds ratios would be less than 0.67 and another 10% would be above 1.01.

**EXERCISE 10.10.** Run the following code for the ICU study and verify the results given in Example 10.2.3. Obtain a plot of the predictive distribution of “true” odds ratios. Obtain a posterior estimate of the proportion of odds ratios that will be below 0.9, 0.8, and 0.6, respectively, and above 1, 1.1, and 1.5, respectively.

```
model{
  for(i in 1:m){
    LOR[i] ~ dnorm(eta[i], tau[i])
    eta[i] ~ dnorm(mu, taub)
    tau[i] <- 1/pow(Se[i],2)
  }
  mu ~ dnorm(0, 0.001)
  taub <- 1/(sigmab*sigmab)
  sigmab ~ dunif(0,100)
  # Median of OR distribution
  medianOR <- exp(mu)
  prob <- step(medianOR - 1)
  # 10th and 90th percentiles of distribution of ORs
  ninetypct <- exp(mu + 1.28*sigmab)
  tenpct <- exp(mu - 1.28*sigmab)
  # Predictive density of ORs
  muf ~ dnorm(mu, taub)
  ORf <- exp(muf)
  for(i in 1:m){ OR[i] <- exp(eta[i]) }
}
list(m=14,
```

```
LOR=c(-0.79493,-0.72392,-0.63227,-0.45604,
      -0.40547,-0.3071,-0.30157,-0.28041,
      -0.15861,0.028988,0.068993,0.100155,
      0.145954,0.200933),
Se=c(0.436627,0.384353,0.332176,0.431258,
     0.629153,0.246629,0.302944,0.29574,
     0.444878,0.448152,0.440058,0.277911,
     0.241851,0.240026)
list(mu=0,sigmab=1)
```

### 10.2.2 Random Slopes and Random Intercepts

Consider longitudinal data in which  $x'_{ij} = (1, t_{ij}, x'_{*ij})$  so that the standard fixed effects regression model includes a time trend

$$\begin{aligned} y_{ij} &= \beta_1 + \beta_2 t_{ij} + x'_{*ij} \alpha + \varepsilon_{ij} \\ &= (1, t_{ij}) \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + x'_{*ij} \alpha + \varepsilon_{ij} \\ &= (1, t_{ij}, x'_{*ij}) \begin{pmatrix} \beta_1 \\ \beta_2 \\ \alpha \end{pmatrix} + \varepsilon_{ij}. \end{aligned}$$

Here, the regression coefficient vector is  $\beta' = (\beta_1, \beta_2, \alpha')$ . In this case, a useful version of model (1) adds random unit-specific intercepts and slopes:

$$\begin{aligned} y_{ij} &= (1, t_{ij}) \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + x'_{*ij} \alpha + (1, t_{ij}) \begin{pmatrix} b_{1i} \\ b_{2i} \end{pmatrix} + \varepsilon_{ij} \\ &= (\beta_1 + b_{1i}) + (\beta_2 + b_{2i}) t_{ij} + x'_{*ij} \alpha + \varepsilon_{ij}, \end{aligned} \quad (10)$$

where

$$\begin{pmatrix} b_{1i} \\ b_{2i} \end{pmatrix} \mid \Sigma_b(\xi) \stackrel{iid}{\sim} N_2(0, \Sigma_b(\xi)) \perp \!\!\! \perp \varepsilon_{ij} \mid \tau \stackrel{iid}{\sim} N(0, 1/\tau).$$

We could simplify the model and assume that the random intercepts and slopes are independent with their own unknown precisions, say  $\tau_{b1}$  and  $\tau_{b2}$ , respectively. In this case  $\xi = (\tau_{b1}, \tau_{b2})'$ .

Alternatively, we can place a prior on the *unstructured* matrix  $\Sigma_b(\xi) \equiv \Sigma_b$ . The *precision matrix* is defined as  $T_b \equiv \Sigma_b^{-1}$  and a standard prior is  $\Sigma_b^{-1} \sim \text{Wishart}(v, B)$ . (A precision matrix does not necessarily contain the scalar precision parameters.) The Wishart distribution generalizes the scaled  $\chi^2$  distribution, see Christensen (2001a, Subsection 1.2.1). The degrees of freedom are  $v$  and  $B$  is a positive definite matrix. The prior expectation is  $vB$ . A common choice for a reference prior is to select  $v$  to be small, and to let  $B = \varepsilon I_2$  with small  $\varepsilon$ . While this prior has expectation zero off the diagonal, the prior allows the intercept and slope random effects to be correlated.

For future reference, we catalogue some code for simulating from the bivariate normal distribution,  $b_i \sim N_2(0, \Sigma_b)$ , and the  $2 \times 2$  Wishart distribution,  $T_b \sim \text{Wishart}(v, B)$ . The particular Wishart distribution below has three degrees of freedom and expectation  $0.003I_2$ . Extension to a higher dimension, say  $k$ , simply involves replacing 2 with  $k$ , and giving a  $k \times k$  matrix  $B$  and a  $k \times 1$  vector for  $m$  in the “list” statement. Note that WinBUGS expects the precision matrix and not the covariance matrix in the specification of the multivariate normal distribution, and it lists  $B$  before  $v$  in the Wishart.

```
model{
  for(i in 1:a){
    b[i,1:2] ~ dmvnorm(m[1:2], Tb[1:2,1:2])
```

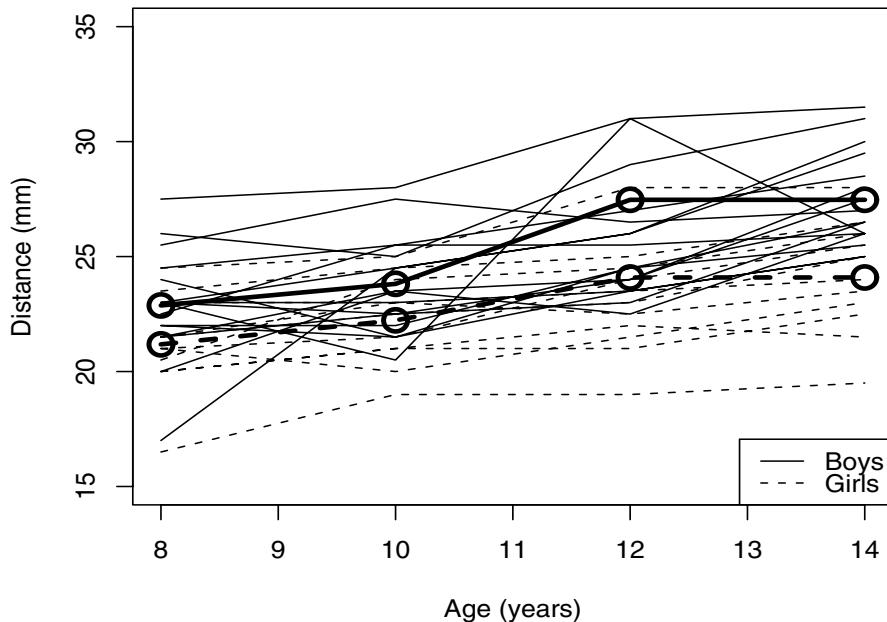


Figure 10.5: Plot of the dental data with sample means for boys and girls.

```

}
Tb[1:2,1:2] ~ dwish(B[1:2,1:2], 3)
}
list(a=, m=c())
B[,1] B[,2]
0.001 0
0      0.001
END

```

**EXAMPLE 10.2.4. Dental Data.** Potthoff and Roy (1964) provide a classic example of longitudinal data. Measurements were obtained from X-rays on the distance (in millimeters) from the pituitary gland to the pterygomaxillary fissure for 16 boys and 11 girls at ages 8, 10, 12, and 14. The covariate Sex is coded as 1 = Male and 0 = Female. We compare boys and girls in both absolute difference and rate of growth.

The data are plotted in Figure 10.5 along with the sample means as a function of Age for boys and girls. We connect data points both for individuals and for overall averages. The resulting curves don't always look like lines. The curves based on averages look parallel, but there may be different slopes between ages 8–10, 10–12 and 12–14. Nonetheless, for our first analysis of these data, we propose a simple model that has parallel trend lines for boys and girls. Imagine two lines superimposed on Figure 10.5 that, to your eye, best fit the two parallel curves based on the overall means. A Sex effect is reflected by having distinct intercepts and the Age effect is reflected in having a non-zero (common) slope. The intercepts appear to be different and the slopes appear positive.

Our analysis based on this model is presented below. In Exercise 10.11, we further explore the data by introducing an interaction between Sex and Age into the model, which allows for different average slopes for males and for females. Then in Exercise 10.18, we explore a model that allows

Table 10.4: Posterior estimates from the dental data model with equal slopes.

	Median	2.5%	97.5%
$\beta_1$	15.9	14.2	17.6
$\beta_2$	0.66	0.52	0.80
$\beta_3$	1.35	-0.28	2.81
$\sigma$	2.09	1.55	2.91

for different but parallel slopes over the different age ranges. The analysis of trend curves is called *profile analysis*.

For the dental data consider the following mixed model:

$$y_{ij} = \mu_{ij} + \varepsilon_{ij}, \quad \varepsilon_{ij} | \tau \stackrel{iid}{\sim} N(0, 1/\tau),$$

$$\mu_{ij} = (\beta_1 + b_{1i}) + (\beta_2 + b_{2i})\text{Age}_{ij} + \beta_3\text{Male}_i + \beta_4\text{Male}_i \times \text{Age}_{ij},$$

$i = 1, \dots, 27 = 16 + 11; j = 1, 2, 3, 4$ . This model allows for different average intercepts for Females and Males ( $\beta_1$  and  $\beta_1 + \beta_3$ , respectively), and also allows for different average slopes ( $\beta_2$  and  $\beta_2 + \beta_4$  for Females and Males, respectively). These are average intercepts and slopes since  $E(b_{ki}) = 0, k = 1, 2$ , under our assumptions. Our initial analysis assumes  $\beta_4 = 0$  and uses independent  $N(0, 10000)$  priors for the other regression coefficients, with  $\tau \sim \text{Gamma}(0.0001, 0.0001)$ . A Wishart( $2, 0.001I_2$ ) prior was placed on the unstructured precision matrix  $T_b = \Sigma_b^{-1}$ , thus our prior expectation for the precision matrix is  $0.002I_2$ .

Posterior median estimates for each child's (random) trajectory,  $\mu_i = (\mu_{i1}, \dots, \mu_{i4})'$ , are presented in the trellis plot in Figure , and posterior estimates for  $\beta$  and  $\sigma = 1/\sqrt{\tau}$  are given in Table 10.4. From this model one might conclude that Age was an important factor and that Sex was not, since the 95% PI for  $\beta_2$  excludes zero and the 95% PI for  $\beta_3$  doesn't. The most intuitive inference for profiles would plot the lines  $15.9 + 0.66\text{Age}$  and  $17.25 + 0.66\text{Age}$  over the age range 8 to 14 for Females and Males, respectively. The slope is nonzero while the difference in intercepts is inconclusive. In Exercise 10.11, we will find that a better analysis results in noticeably different conclusions due to a statistically important interaction between Age and Sex, making the slope for Males different than the slope for Females. These data are also analyzed in Example 10.4.1 and Exercises 10.17 and 10.18.

**EXERCISE 10.11. Dental Data.** (a) Graph the regression lines  $15.9 + 0.66\text{Age}$  and  $17.25 + 0.66\text{Age}$  based on the model with no interaction. Superimpose on this plot the average curves that are given in Figure 10.5. How well did your mind's eye create these lines? (b) Write WinBUGS code to analyze the dental data using the mixed model with random intercepts and equal slopes. Use the diffuse Wishart distribution from Example 10.2.4 to jointly model the random intercept and slope combinations for each individual. (See the WinBUGS code just above Example 10.2.4.) Verify that the estimates given in Table 10.4 are correct up to Monte Carlo error. Also obtain the DIC value for this model. (c) Now expand the model to include the interaction term between Sex and Age. (i) Analyze the data. As part of the analysis, use the estimates of the regression parameters to obtain a plot like the one in (a). Comment. (ii) Calculate DIC for this model. Make inferences for  $\beta_4$ . Decide whether the interaction term belongs in the model. (d) Perform a sensitivity analysis using your selected model. Make complete inferences for your final selected model. (e) Reanalyze the data assuming independence between the intercept and slope random effects using independent uniform or gamma priors for their precisions. Compare the analysis with that of part (c).

**EXERCISE 10.12.** (a) Derive formulas for the variance of  $y_{ij}$  and the covariance between  $y_{ij}$  and  $y_{ik}$  under the random intercepts and slopes model (10) with a  $2 \times 2$  unstructured covariance matrix

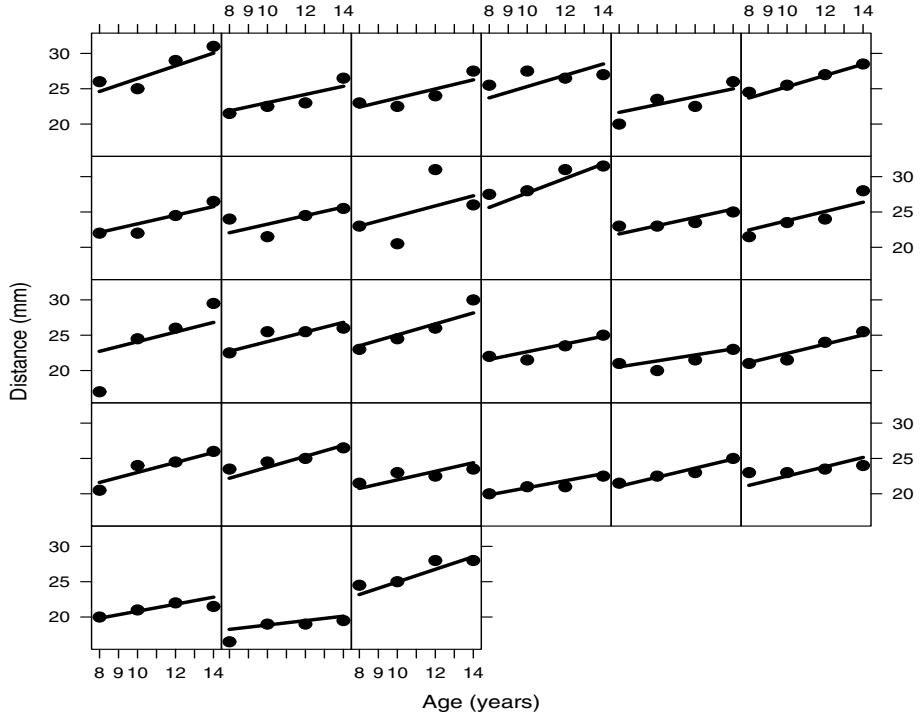


Figure 10.6: Plots of the fitted (posterior median) random intercept and slope mixed model to each child's data.

$\Sigma(\xi) \equiv \Sigma$ . (b) What are  $X_i$  and  $Z_i$  for model (10)? (c) What is the marginal distribution of the data based on model (10)?

### 10.3 Multivariate Normal Models

We begin with a random sample of balanced data  $y'_i = (y_{i1}, \dots, y_{ik})'$ ,  $i = 1, \dots, n$ , where  $y_{ij}$  denotes the  $j$ th of  $k$  measurements taken on unit  $i$ . (In comparison with the previous section, here  $n$  is the number of units rather than  $a$  and  $k$  is the common number of observations per unit, rather than having  $n_i$  observations on unit  $i$ . The total number of observations is  $nk$  rather than  $n$  as in the previous section.) The data are assumed to be independent across units ( $y_i \perp\!\!\!\perp y_\ell, i \neq \ell$ ), but possibly correlated within units, i.e.,  $\text{Cov}(y_{is}, y_{it})$  is not necessarily 0. The most general one-sample multivariate normal model assumes

$$y_i | \mu, \Sigma \stackrel{iid}{\sim} N_k(\mu, \Sigma), \quad (1)$$

where  $\mu = (\mu_1, \dots, \mu_k)'$  and  $\Sigma$  is an arbitrary  $k \times k$  positive definite covariance matrix with  $s, t$  element  $\sigma_{st} = \sigma_{ts}$ ,

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1,k-1} & \sigma_{1k} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2,k-1} & \sigma_{2k} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \sigma_{k-1,1} & \sigma_{k-1,2} & \cdots & \sigma_{k-1,k-1} & \sigma_{k-1,k} \\ \sigma_{k1} & \sigma_{k2} & \cdots & \sigma_{k,k-1} & \sigma_{kk} \end{pmatrix}. \quad (2)$$

There are  $(k+1)k/2$  distinct parameters in  $\Sigma$ , so for  $k$  large, a fair amount of data are needed to adequately estimate the covariance matrix. The corresponding correlation matrix is  $R = \{\rho_{st}\}$

where  $\rho_{st} = \text{Corr}(y_{is}, y_{it} | \mu, \Sigma) = \sigma_{st} / \sqrt{\sigma_{ss} \sigma_{tt}}$ . In the next subsection we discuss various models for the covariance matrix.

### 10.3.1 Parameterized Covariance Matrices

If our variables are age, blood pressures, cholesterol, weight, and height, it is difficult to imagine any structure that would apply to the covariances, so the *unstructured covariance* of (2) would be used. The unstructured covariance model can be used for any multivariate data, e.g., repeated cholesterol measurements on the same individual; however, repeated measures suggest alternative *structured covariance models*.

A key feature of structured covariance models is that the variables  $y_{ij}$  are all measured on the same scale, e.g., repeated cholesterol measurements. In general, a structured covariance matrix  $\Sigma$  depends on a parameter vector  $\theta$ , i.e.,

$$\Sigma \equiv \Sigma(\theta).$$

The specific covariance models detailed in this subsection assume constant variance ( $\sigma_{ss} \equiv \sigma^2$  for all  $s$ ), so the models can all be written as  $\sigma^2$  times a correlation model, say,

$$\Sigma(\theta) = \sigma^2 R(\rho)$$

where the correlation matrix  $R$  depends on a vector of parameters  $\rho$ , thus  $\theta' = (\sigma^2, \rho')$ . The dimension of  $\rho$  varies from 0 to  $k - 1$ . See Wolfinger (1996) for a discussion of models with heterogeneous variances.

Unstructured covariance is the most complicated model we can assume. The simplest model we can assume is independence with equal variances, i.e.,

$$\Sigma = \sigma^2 I_k.$$

These constitute the two extremes, one with the maximum number of covariance parameters and the other with only one parameter.

Suppose the data in the vector  $y_i$  are from a random sample of  $k$  patients taken from the  $i$ th hospital in a random sample of hospitals. Because we have a random sample of hospitals, it may be reasonable to assume that the  $y_i$ s have a common covariance matrix  $\Sigma$ . But because the patients within a hospital are a random sample, it is reasonable to assume that they are exchangeable. In particular, the patient variables  $y_{ij}$  should have the same variance and any two of them should have the same correlation. This implies that the correlation matrix has a diagonal of 1s and everything else is a scalar constant  $\rho$ , hence

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho & \rho \\ \rho & 1 & \cdots & \rho & \rho \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho & \rho & \cdots & 1 & \rho \\ \rho & \rho & \cdots & \rho & 1 \end{pmatrix}.$$

This is called *compound symmetry* (CS) or *intraclass correlation* or *equicorrelation*. To make the covariance matrix positive definite, we must have  $-1/(k-1) < \rho < 1$ . The random intercept models of the previous section determined CS covariance structures with positive correlation but they allowed for unbalanced data. *Compound symmetry is appropriate when the random variables under consideration are exchangeable.*

In balanced longitudinal studies, the observation  $y_{ij}$  is observed at time  $t_j$ . Our remaining models assume equal time spacing, i.e.,  $t_{j+1} - t_j = d$  for some  $d$  and all  $j$ . With equal spacing, assume a constant lag  $s$  correlation, say,

$$\rho_s \equiv \text{Corr}(y_{ij}, y_{i,j+s}),$$

computed between any measurements collected at times separated by  $|t_{j+s} - t_j| = sd$  units. With a constant variance, this results in a *banded* covariance model,

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-3} & \rho_{k-2} & \rho_{k-1} \\ \rho_1 & 1 & \rho_1 & \ddots & \rho_{k-4} & \rho_{k-3} & \rho_{k-2} \\ \rho_2 & \rho_1 & 1 & \ddots & \ddots & \rho_{k-4} & \rho_{k-3} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \rho_{k-3} & \rho_{k-4} & \ddots & \ddots & 1 & \rho_1 & \rho_2 \\ \rho_{k-2} & \rho_{k-3} & \rho_{k-4} & \ddots & \rho_1 & 1 & \rho_1 \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & \rho_2 & \rho_1 & 1 \end{pmatrix},$$

where now,  $\theta = (\sigma^2, \rho_1, \rho_2, \dots, \rho_{k-1})'$ . These are positive definite *Toeplitz* matrices. Our remaining models are special cases of this. Ideally, we would put a prior on the parameters of the covariance matrix that ensures that the covariance matrix is positive definite but it is by no means easy to determine what parameters in a Toeplitz matrix will make it positive definite, so the support of such prior distributions is not always easy to find.

A popular model is the first order *autoregressive* or AR(1) model which assumes that correlation decays exponentially with increasing lags in time, i.e.,  $\rho_s = \rho^s$  with  $-1 < \rho < 1$ . For  $k = 4$ ,

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}.$$

If  $\rho = 0.8$ , observations that are two lags apart have correlation 0.64, and if they are three lags apart, the correlation is 0.512.

Another time series based model is the first order *moving average* or MA(1) model. It assumes that the only non-zero correlation occurs at lag 1. For  $k = 4$ ,

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho & 0 & 0 \\ \rho & 1 & \rho & 0 \\ 0 & \rho & 1 & \rho \\ 0 & 0 & \rho & 1 \end{pmatrix}.$$

Here  $-0.5 < \rho < 0.5$ . Combining the AR(1) and MA(1) models leads to an *autoregressive moving average* model, specifically an ARMA(1,1), that has lag  $s$  correlation  $\rho_s = \gamma\phi^{s-1}$ . For  $k = 4$ ,

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \gamma & \gamma\phi & \gamma\phi^2 \\ \gamma & 1 & \gamma & \gamma\phi \\ \gamma\phi & \gamma & 1 & \gamma \\ \gamma\phi^2 & \gamma\phi & \gamma & 1 \end{pmatrix}.$$

Frequently, any parameter values with  $-1 < \gamma < 1$  and  $-1 < \phi < 1$  are allowed but some such parameters do not determine a positive definite matrix. From time series theory, the simplest way around this problem is to allow  $-1 < \theta < 1$  and  $-1 < \phi < 1$  and set

$$\gamma = \frac{(\phi - \theta)(1 - \phi\theta)}{(1 - \phi^2) + (\phi - \theta)^2}.$$

We will not discuss general AR( $p$ ), MA( $q$ ), or ARMA( $p, q$ ) models that are primarily used to analyze a single, long time series of data, see Shumway (1988), Shumway and Stoffer (2006), or Christensen (2001a, Chapter. 5). They can be used to generate more general positive definite

Toeplitz covariance matrices, although the constraints on the parameters that provide positive definite matrices become more complicated.

We presumed balanced data with  $k$  observations on each unit in the sample. For longitudinal data, we assumed equal spacing. In longitudinal data or two stage sampling, observation vectors frequently have different lengths or unequal spacing, so additional issues arise in modeling covariance structures. For example, if correlations depend on the time between measurements and those differ for two individuals, it makes no sense to model their correlations as being equal. The form of the correlation matrix should be different for such individuals. In particular, a generalization of the AR(1) model for unequally spaced time intervals is the *exponential covariance function*

$$\text{Cov}(y_{ij}, y_{ik}) = \sigma^2 e^{-\theta_1 |t_j - t_k|}.$$

This model only provides positive correlations and requires  $\theta_1 > 0$ . See Section 10.4 for an alternate approach.

**EXERCISE 10.13.** *Patterned Covariance Matrices.* (a) Suppose you have repeated measurements at times 1, 2, 3, 4. You believe observations should be exponentially less correlated the further apart they are in time. Give an appropriate covariance structure. (b) Repeat (a) if the times were 1, 3, 5, 7. Is there any important distinction between this model and the one proposed in (a)? Explain. (c) Repeat (b) if the times were 1, 3, 7, 15. (d) Repeat (b) if the times were 1, 3, 4, 9.

Prior specifications for covariance matrices are difficult. Not only must variances be positive and correlations restricted to  $(-1, 1)$  but a covariance matrix must be positive definite. While there are methods available in the literature for handling various covariance structures, we stick to a few relatively simple cases.

Diffuse prior specifications on  $\sigma$ ,  $\sigma^2$ , or  $\tau$  include using a gamma distribution with mean 1 and large variance or a uniform distribution over a broad range. “Broad” means not excluding values that are remotely possible. On the other hand, it is sensible to exclude values that are impossible. For example, the standard deviation of women’s heights certainly cannot be larger than 10 feet. Otherwise, we would see women more than 20 feet tall. We could restrict the standard deviation to be smaller but 10 feet is probably a sufficient restriction.

For  $\rho$  in the CS model, a  $U(-1/(k-1), 1)$  provides an appropriate diffuse prior. For the AR(1) and MA(1) models, the  $U(-1, 1)$  and  $U(-0.5, 0.5)$  distributions on  $\rho$  are appropriate, respectively. An alternative choice for the MA(1) based on time series theory has  $\theta \sim U(-1, 1)$  with  $\rho = \theta/(1 + \theta^2)$ . Independent  $U(-1, 1)$ s can be used for  $\phi$  and  $\theta$  in the ARMA(1,1). If the parameters are known to be positive, Beta priors can be used for  $\rho$ ,  $\phi$ , and  $\theta$ . These allow substantive prior information. Similarly, generalized Beta priors defined on arbitrary intervals can be used for the entire allowable range of the parameters. With an exponential covariance function, the positive parameter  $\theta_1$  can have a diffuse Gamma prior. In our experience, correlations in grouped data are typically positive.

An unstructured covariance matrix,  $\Sigma$ , is simply an arbitrary covariance matrix as presented in (2). A Wishart distribution is commonly used as a prior on the precision matrix  $\Sigma^{-1}$ . This *inverse Wishart* distribution (or even a Wishart) with a positive definite parameter matrix when used as the prior for  $\Sigma$  ensures that the covariance matrices are positive definite with probability one. Again, the precision matrix does not necessarily contain any of the individual scalar precisions.

Uncertainty about the mean vector  $\mu$  can be modeled with a multivariate normal. More simply, independent univariate normals can be used for each  $\mu_j$ .

**EXAMPLE 10.3.1** *Reference Values for IL-1 $\beta$ .* The enrolled subjects (units) were healthy and had no relevant health changes during the study, so assuming that all observations have the same mean  $\mu$  and variance  $\sigma^2$  over time seems reasonable. The sampling model is taken as

$$y_i | \mu, \Sigma \stackrel{iid}{\sim} N_6(\mu J_6, \Sigma(\theta))$$

where  $J_6 \equiv (1, 1, 1, 1, 1, 1)'$ . We considered all six models: independence, compound symmetry, AR(1), MA(1), ARMA(1,1) and unstructured. A  $N(0, 1000)$  prior was used for  $\mu$ , a  $U(0, 100)$  prior was used for the standard deviation, and a uniform prior distribution over the appropriate parameter space was placed on  $\rho$ . In the ARMA(1,1) we took  $\theta$  and  $\phi$  as independent  $U(-1, 1)$ s, which induces a prior on  $\gamma$ . For the unstructured covariance model, we used  $\Sigma^{-1} \sim \text{Wishart}(6, 0.01I)$ .

Comparing the models with different covariance structures using DIC statistics gives

Cov	Indep.	CS	AR(1)	MA(1)	ARMA(1,1)	Unstruct.
DIC	580	533	538	555	529	539

Among these choices the ARMA(1,1) is preferred. Although ARMA(1,1) fits best, CS and AR(1) are not bad. Both outperform the unstructured covariance matrix, and anything that outperforms unstructured cannot be too bad. Since CS is not too bad, the mixed model is not too bad. We obtained the predictive distribution of a single scalar future observation from the ARMA(1,1) covariance model and calculated a point prediction (median of the predictive distribution) of 4.9 with corresponding 95% prediction interval for  $y_f$  of (0.45, 55.2). Recall from Table 10.1 that our point prediction and interval based on the mixed model (resulting in a CS covariance structure) were 4.7 (0.39, 56.5). The inferences are so similar that we doubt anyone would care about the differences. The huge asymmetries of the PIs about the median are due to transforming results obtained on the log-scale back onto the original scale.

The following WinBUGS code for the CS model is hybridized from our mixed model code of Example 10.2.1. The key difference is that it incorporates the multivariate normal distribution `dmmnorm`, which WinBUGS parameterizes in terms of the precision matrix.

```

model{
  for(i in 1:30){
    for(j in 1:6){
      logy[i,j] <- log(y[i,j])
    }
  }
  for(i in 1:30){logy[i,1:6] ~ dmmnorm(m[1:6],precision[1:6,1:6])}
  for(j in 1:6){
    for(k in 1:6){
      covariance[j,k] <- sigma2*pow(rho, step(abs(j-k)-0.5))
    }
  }
  for(i in 1:6){ m[i] <- mu }
  precision[1:6,1:6] <- inverse(covariance[1:6,1:6])
  sigma ~ dunif(0,100)
  mu ~ dnorm(0,0.001)
  L <- -1/(6-1)
  rho ~ dunif(L,1)
  sigma2 <- sigma*sigma
  tau <- 1/sigma2
}
y[,1] y[,2] y[,3] y[,4] y[,5] y[,6]
22.8 17.6 11.6 7 8.78 6.6
3.14 14 7.13 3.66 1.992 7.46
12.7 5.63 6.4 4.14 5.3 0.784
remaining 27 data lines go here
END

```

Example 10.2.1 gives virtually the same estimates as obtained from this iid multivariate normal analysis with compound symmetry, which is not surprising since the models for the data are identical. However, the random effects model runs substantially faster in WinBUGS. That is because, using the current code, WinBUGS must take the time to numerically invert the  $6 \times 6$  covariance matrix at every iteration of the Gibbs sampler. Below, we give explicit formulas for the inverse of the CS and AR(1) covariance matrices. Incorporating these forms makes the code run considerably faster.

#### 10.3.1.1 Analytic Formulas for CS and AR(1) Precision Matrices

For  $\Sigma \equiv \{\sigma_{st} : s, t = 1, \dots, k\}$ , let  $T = \{\tau_{st}\}$  denote the corresponding precision matrix  $\Sigma^{-1}$ . The CS covariance matrix has  $\sigma_{ss} = \sigma^2$  and  $\sigma_{st} = \sigma^2\rho$ ,  $s \neq t$ . The analytical formula for the CS precision matrix is

$$\tau_{ss} = \frac{1}{\sigma^2(1-\rho)} \left\{ 1 - \frac{\rho}{1+(k-1)\rho} \right\}, \quad \tau_{st} = \frac{-1}{\sigma^2(1-\rho)} \left\{ \frac{\rho}{1+(k-1)\rho} \right\}.$$

The AR(1) covariance matrix has  $\sigma_{st} = \sigma^2\rho^{|s-t|}$ . The precision matrix has

$$\begin{aligned} \tau_{11} = \tau_{kk} &= \frac{1}{\sigma^2(1-\rho^2)}; & \tau_{ss} &= \frac{1}{\sigma^2(1-\rho^2)}(1+\rho^2), \quad s = 2, \dots, k-1; \\ \tau_{st} &= \frac{-\rho}{\sigma^2(1-\rho^2)}, \quad |t-s|=1; & \tau_{st} &= 0, \quad |t-s| \notin \{0, 1\}. \end{aligned}$$

**EXERCISE 10.14.** (a) Run the code given earlier to get a feeling for the speed of iterations when WinBUGS inverts the covariance matrix at each iteration. (b) Modify the code to use the analytical form for the CS precision matrix. See how much faster it runs than the original code. Compare final results with those in (a). (c) Modify the code to incorporate the AR(1) precision matrix in its analytical form. Also modify the code to handle the unstructured covariance matrix. Obtain DIC statistics to compare models with constant mean structure and (i) CS, (ii) AR(1), and (iii) unstructured covariance models. (d) Repeat (c) only now model a different mean at each time point and compute the DIC statistic. (e) Compare all six models from (d) and (c) and decide if changing means over time is needed. If means change, you might consider modeling the means as lines or quadratics in time.

**EXERCISE 10.15.** Davis (2002, p. 16) reproduced data from Deal et al. (1979) on ventilation volumes ( $\text{min}^{-1}$ ) measured on eight subjects at six different temperatures. Analyze the data (available at our website) to test whether temperature affects ventilation volume. Plot the data and lines connecting the sample means, the posterior medians, and the 95% intervals at each temperature. Give conclusions.

## 10.4 Multivariate Normal Regression

We now generalize Section 3 to allow the components of  $\mu$  to depend on covariates and allow unbalanced longitudinal data. The models are called *multivariate normal regression* or *growth curve* models.

We treat unbalanced data as balanced data with missing observations. Consider *all possible* observation times across all units and denote them  $t_j$ ,  $j = 1, \dots, k$ . Define  $\Sigma(\theta)$  as the  $k \times k$  covariance matrix appropriate if all units were observed at all times. Any of the parameterized covariance matrices described in the previous section can be used to model  $\Sigma(\theta)$ , although all of those models except compound symmetry and the exponential covariance function require equal spacing. For

unit  $i$  we observe  $y_i \equiv (y_{i1}, y_{i2}, \dots, y_{ik_i})'$  with  $k_i \leq k$  and observations taken at times  $t_{ij}$ , which constitute a subset of the possible times  $t_j$ ,  $j = 1, \dots, k$ . The covariance matrix for  $y_i$  is  $\Sigma_i(\theta)$  and is the submatrix of  $\Sigma(\theta)$  that corresponds to the times  $t_{i1}, \dots, t_{ik_i}$ .

The multivariate regression model is

$$y_i | \beta, \theta \stackrel{ind}{\sim} N_{k_i}(X_i \beta, \Sigma_i(\theta)), \quad (1)$$

where  $X_i$  is a known fixed  $k_i \times r$  matrix. Since  $\beta$  is the same for each unit  $i$ , the predictor variables must be the same for every time and every unit. Each column of  $X_i$  contains values for an individual predictor. In longitudinal data analysis, the  $(j, s)$ th element of  $X_i$ , say  $x_{ijs}$ , denotes the value of covariate  $s$  measured at time  $t_{ij}$  for unit  $i$ . These can be either fixed over time (e.g., sex) or time varying (e.g., blood pressure or  $t_{ij}$  or some function of time,  $f(t_{ij})$ ). For studies conducted over a relatively short period of time, a baseline age is often used as a fixed-time covariate. An intercept term is included by setting the first column of each  $X_i$  to one. The rank of  $X_i$  is often much less than  $r$  because any covariates that do not change with time are linearly dependent within  $X_i$ . Alternatively, we could write the model

$$y_i \stackrel{ind}{\sim} N_{k_i}(\mu_i, \Sigma_i(\theta)), \quad \mu_i = X_i \beta.$$

The model (10.3.1) can be generalized to unbalanced data as a special case of model (1). Taking  $\beta = (\mu_1, \dots, \mu_k)'$ , choose  $X_i$  to be the  $k_i \times k$  submatrix of  $I_k$  with rows that correspond to the times  $t_{i1}, \dots, t_{ik_i}$ .

**EXERCISE 10.16.** (a) Give the form of the matrix  $X_i$  when longitudinal data are collected at five time points, the expected responses in time follow a linear trend, and the intercept depends on sex and baseline age. (b) Modify  $X_i$  to allow different time trends for men and women.

**EXAMPLE 10.4.1.** *Dental Data.* Example 10.2.4 introduced the growth curve data of Potthoff and Roy (1964) that were collected over several years. The data are balanced and equally spaced with four times (Ages) per child. Our regression model involves two straight lines over Ages, one for boys and one for girls. The key feature is that the four observations taken at different ages on each child are dependent. The  $j$ th row of  $X_i$  is taken as

$$x'_{ij} = (1, \text{Age}_j, \text{Male}_i, \text{Age}_j \times \text{Male}_i),$$

so  $\beta_1$  is the intercept for girls,  $\beta_2$  is the slope for girls,  $\beta_1 + \beta_3$  is the intercept for boys, and  $\beta_2 + \beta_4$  is the slope for boys. Although Ages could be different for each child, in these data they are not. We also assume *equal* and *unstructured* covariances for boys and girls with the prior for the unstructured precision matrix  $\Sigma^{-1}$  taken as a diffuse Wishart(4, 0.001 $I_4$ ) distribution. Independent  $N(0, 1000)$  priors were used for the regression coefficients.

The growth rate for boys,  $\beta_2 + \beta_4$ , has posterior median and 95% probability interval 0.83, (0.66, 0.99). For girls, the growth rate is  $\beta_2$  with posterior values 0.48, (0.28, 0.67). The intervals have almost no overlap. The difference in growth rates,  $\beta_4$ , has posterior values 0.35, (0.09, 0.61). The posterior probability that the growth rate for boys exceeds that for girls is  $\Pr(\beta_4 > 0 | Y) = 1$ . The estimated mean distances and differences for boys compared to girls are presented in Table . The posterior probabilities that the mean distance for boys exceeds the mean for girls ( $\Pr[\beta_3 + \beta_4 \text{Age} > 0 | Y]$ ) at ages 8, 10, 12, and 14 are 0.94, 0.99, 1.00, and 1.00, respectively.

**EXERCISE 10.17.** *Dental Data.* (a) Calculate the DIC statistic for the model that has separate linear trends for boys and girls using a common *compound symmetric* covariance structure. (b) Calculate DIC for the model that has separate linear trends and *different compound symmetric* covariance matrices for boys and for girls. (c) Select a preferred model from among those considered in (b), and the model with *equal and unstructured covariances*, like the one that was considered

Table 10.5: Posterior estimates for mean distance and the difference in mean distance for boys and girls in the dental data.

Units	Mean/Median	SD	(95% PI)
Boys, age 8	22.46	0.53	(21.42, 23.49)
Boys, age 10	24.11	0.49	(23.15, 25.08)
Boys, age 12	25.76	0.51	(24.76, 26.77)
Boys, age 14	27.42	0.58	(26.27, 28.56)
Girls, age 8	21.23	0.60	(20.04, 22.42)
Girls, age 10	22.19	0.56	(21.08, 23.28)
Girls, age 12	23.14	0.58	(22.00, 24.28)
Girls, age 14	24.09	0.66	(22.79, 25.39)
B–G, age 8	1.22	0.80	(−0.36, 2.81)
B–G, age 10	1.92	0.74	(0.46, 3.39)
B–G, age 12	2.62	0.77	(1.10, 4.15)
B–G, age 14	3.32	0.88	(1.59, 5.06)

in Example 10.4.1. Present inferences from your chosen model. (d) Compare inferences with those from your final model in Exercise 10.11.

**EXERCISE 10.18.** *Dental Data.* This exercise develops a model for the dental data with non-linear parallel profiles. Let  $\mu_{ij} : i = 1, 2, j = 1, 2, 3, 4$  be the collection of mean responses on individuals with Sex  $i$  and Age  $j$ . Here  $i = 1$  corresponds to males and  $j = 1$  corresponds to 8 year-olds. The constraint on the model is that  $\mu_{1j} - \mu_{2j} = \delta$  for some constant  $\delta$ , so there is a constant difference in the effect of Sex, regardless of Age. The mean structure for this model has no interaction which implies parallel curves (not restricted to be lines). There are two goals in this analysis. The first is to make inferences about  $\delta$ , the common effect of Sex. The second is to assess the effect of age, only now there isn't a single slope for each sex with which to describe this effect. Consider at least two distinct covariance structures and select one of those using DIC. Finally, compare your final inferences with those obtained in previous exercises and with those from Example 10.2.4. Write up the analysis as a report with an introduction, development of the model, analysis of the data, and final conclusions. There are multiple correct ways to do the analysis.

## 10.5 Posterior Sampling and Missing Data

For the random intercepts model (10.2.4) with independent multivariate normal, scalar normal, gamma, and gamma distributions for  $\beta$ , the  $b_i$ s,  $\tau_b$ , and  $\tau$ , respectively, the full conditional  $p(\beta | b_i s, \tau_b, \tau, Y)$  is multivariate normal, the full conditionals  $p(b_i | \beta, \tau_b, \tau, Y)$  are independent normal distributions for each  $b_i$ ,  $p(\tau | \beta, \tau_b, b_i s, Y)$  is gamma, as is  $p(\tau_b | \beta, \tau, b_i s, Y)$ .

**EXERCISE 10.19.** Using the same prior as in Exercise 10.6, derive the full conditionals for model (10.2.6). The derivations are much the same as for model (10.2.5), cf. Exercise 10.6, but with additional algebraic complexity.

Posterior sampling is more difficult for the general mixed model (10.2.3) and the multivariate regression model (10.4.1), including its special case model (10.3.1). For the general mixed model (10.2.3),  $\theta$  denotes  $(\sigma, \xi')'$ . With a multivariate normal prior for  $\beta$ , the full conditional  $p(\beta | \theta, Y)$  is multivariate normal, while the full conditional for  $p(\theta | \beta, Y)$  will generally be unrecognizable. Something like the Metropolis algorithm is needed to sample  $\theta$ . A better approach for the general mixed model with an independent gamma prior on  $\tau$  incorporates the random effects. The

full conditionals are the multivariate normal  $p(\beta | b, \xi, \tau, Y)$ , the independent multivariate normals  $p(b_i | \beta, \xi, \tau, Y)$ , the gamma  $p(\tau | \beta, \xi, b, Y)$ , and the probably unrecognizable  $p(\xi | \beta, \tau, b, Y)$ .

In dealing with unbalanced data, specifying all of the matrices  $\Sigma_i(\xi)$  for model (10.4.1) can be quite painful. If the data are nearly balanced, one can treat the irregularities as missing data. This is accomplished by modeling the missing observations as if they had been observed. Since they weren't observed, the missing observations will be treated as parameters. The Gibbs sampler performs exactly as with balanced data except it now involves the additional step of obtaining the full conditionals for the missing cases. The full conditionals for the missing data are straightforward. In WinBUGS, after setting up the balanced model, a simple way to handle missing data is to place NA in any data slot that corresponds to missing data. WinBUGS will then automatically sample from the appropriate full conditional distributions. WinBUGS will expect initial values for any missing data. This may be a good time to use the `gen inits` command in WinBUGS. If a large proportion of data points are missing, the code may run very slowly due to the burden of sampling so many extra full conditionals.

---

## Chapter 11

---

# Count Data

---

We now consider regression models for count data, where the counts have no clear upper bounds. Specifically we consider models in which the response has a Poisson distribution or a mixture of Poisson distributions. Our focus is on models with linear structure. Section 1 considers Poisson regression. Section 2 introduces the problem of over-dispersion and uses mixture models including zero-inflated Poisson models to deal with over-dispersion. Section 3 briefly discusses longitudinal Poisson modeling.

### 11.1 Poisson Regression

The basic model for Poisson regression is

$$y_i|\lambda_i \stackrel{ind}{\sim} \text{Pois}(\lambda_i), \\ \log(\lambda_i) = x_i' \beta,$$

$i = 1, \dots, n$ , which is a log-linear model. It is possible, but uncommon, to use functions other than the log.

**EXAMPLE 11.1.1.** Bissell (1972) presents data on the number of faults in rolls of fabric having different lengths. The data are given in Table 11.1. It is reasonable to model the number of faults in any roll of fabric as Poisson and to use the length of the roll as a predictor variable for the number of faults. Thus we assume that the  $n = 32$  observations are independent random variables  $y_i$  with  $y_i|\lambda_i \sim \text{Pois}(\lambda_i)$ .

A reasonable model for the data might be that the expected number of faults  $\lambda_i$  is some number  $\theta$  times the length of the piece of fabric, say  $M_i$ , i.e.,

$$\lambda_i = \theta M_i. \quad (1)$$

Such a model assumes that the faults are being generated at a constant rate, and therefore the expected number of faults is proportional to the length. We can rewrite model (1) as a log-linear model

$$\log(\lambda_i) = \log(\theta) + \log(M_i)$$

or, equivalently,

$$\log(\lambda_i) = \beta_1 + (1) \log(M_i), \quad (2)$$

where  $\beta_1 \equiv \log(\theta)$ . Note that  $\theta$  needs to be a positive number, so an appropriate, indeed the conjugate, prior is the gamma distribution. Alternatively, we might use a log-normal prior distribution on  $\theta$ , i.e., we might use a normal distribution on  $\beta_1$ .

This is really just a one-sample problem with different windows. In Section 5.3 we discussed one-sample problems with the same counting window size, i.e.,  $M = M_1 = \dots = M_n \equiv 1$ , and two-sample problems with different counting windows.

Fitting model (2) with a  $N(0, 1000)$  prior on  $\beta_1$  and fitting model (1) with a prior  $\theta \sim \text{Gamma}(0.001, 0.001)$  give similar posterior results.

Table 11.1: *Textile faults.*

Roll( $i$ )	Length( $M_i$ )	Faults( $y_i$ )	Roll( $i$ )	Length( $M_i$ )	Faults( $y_i$ )
1	551	6	17	543	8
2	651	4	18	842	9
3	832	17	19	905	23
4	375	9	20	542	9
5	715	14	21	522	6
6	868	8	22	122	1
7	271	5	23	657	9
8	630	7	24	170	4
9	491	7	25	738	9
10	372	7	26	371	14
11	645	6	27	735	17
12	441	8	28	749	10
13	895	28	29	495	7
14	458	4	30	716	3
15	642	10	31	952	9
16	492	4	32	417	2

Model	Parameter	mean	sd	2.50%	median	97.50%
(2)	$\beta_1$	-4.195	0.05927	-4.312	-4.194	-4.081
(2)	$\theta = e^{\beta_1}$	0.0151	0.000894	0.01341	0.01508	0.0169
(1)	$\theta$	0.0151	0.000897	0.01339	0.015	0.01692

If we generalize model (2) it looks more familiar as a regression. Using a simple linear regression structure with  $\log(M_i)$  as a predictor variable, we have the more general model

$$\log(\lambda_i) = \beta_1 + \beta_2 \log(M_i). \quad (3)$$

Model (3) indicates that the rate of fabric faults increases (or decreases) with the length of the roll. (Perhaps faults are more or less common near the ends of a roll so that longer rolls have relatively fewer faults.) Model (2) is the special case of model (3) with  $\beta_2 \equiv 1$ , so we would like to compare the models.

Fitting model (3) with independent  $N(0, 1000)$  priors on  $\beta_1$  and  $\beta_2$  gives posterior results:

Model	Parameter	mean	sd	2.50%	median	97.50%
(3)	$\beta_1$	-4.17	1.119	-6.547	-4.115	-2.039
(3)	$\beta_2$	0.9959	0.173	0.6659	0.9876	1.364

Perhaps the most interesting result is how near the posterior mean and median of  $\beta_2$  are to 1, although the 95% PI shows some variability about 1. The deviance information criterion provides a convenient way to compare models (2) and (3). The respective DICs are 189.8 and 191.8, which indicates a preference for model (2).

**EXERCISE 11.1.** Explain in detail how to construct and implement in WinBUGS a BCJ prior for  $\beta$  in Poisson regression. Use model (3) from Example 11.1.1 as an illustration.

**EXERCISE 11.2.** Watkins, Bergman, and Horton (1994) presented the data given in Table 11.2 on a complicated designed experiment that generated counts. The dependent variable is the number of ends cut by a tool. Three observations were made for each experimental condition. The factors in the design are the first five factors listed after Run.

Run	Chaser	Coolant	Speed	Pipe	Rake Angle	Spindle
Run	Ch	Cl	Spd	P	RA	Spn

There are two different chasers, two different coolants, the two speeds were coded as intermediate (0) and high (1), two different pipes, and two different rake angles. (Technically, the experiment was a half replication of a treatment structure with five factors each at two levels, i.e., a half rep. of a  $2^5$  or a  $2^{5-1}$ , see Christensen, 2005, Chapter 17.) In addition, one categorical covariate was observed. On each run it was noted whether the spindle was left (0) or right (1,2,3). It was initially believed that the spindle position would not be important. The three repetitions of run 4 were performed to verify that the spindle position was an important factor. In the course of the experiment, two new heads were installed on the right spindle. A new head was installed prior to run number 8 ( $Spn = 2$ ) and also prior to the second observation of run 15 ( $Spn = 3$ ). As a first pass, it seems reasonable to treat all of the observations as independent Poisson random variables.

All the factors except Spindle are binary, so they are coded as 0/1 and treated in the usual way. For Spindle, we need to create four indicator variables for the four groups. Use all four indicators and delete the intercept from the model. A main effects only ANOVA model is equivalent to fitting a multiple regression on these nine variables,

$$y_{ij} | \lambda_i \stackrel{ind}{\sim} \text{Pois}(\lambda_i), \quad i = 1, \dots, 19, j = 1, 2, 3,$$

$$\log(\lambda_i) = \beta_1 Ch_i + \beta_2 Cl_i + \beta_3 Spd_i + \beta_4 P_i + \beta_5 RA_i + \beta_6 Spn_{1i} + \beta_7 Spn_{2i} + \beta_8 Spn_{3i} + \beta_9 Spn_{4i}.$$

(a) Fit the model using independent  $N(0, 10^3)$  priors for the regression coefficients. Which factors are important for predicting cut counts? (b) Now consider removing terms from the “full” model in part (a), one at a time. After removing a term, run the code and obtain the DIC in each instance. You will have to use each removed term in the code in a benign way so as not to provoke an error message from WinBUGS. Also, the Spn terms must either be all in or all out. Mimicing stepwise regression, remove the variable that corresponds to the largest drop in DIC, unless the DIC values are all larger than for the “full” model, in which case you select the full model. If you delete a variable, then repeat the process to see if you can delete a second variable, and continue until you decide to keep all the remaining terms. Finally, for the selected model, obtain estimated counts for all combinations of remaining terms in the model and comment on the differences. (c) A model with all two factor interactions can be constructed by removing one of the  $Spn_{ki}$  variables, adding new predictor variables that are the pairwise products between all of the remaining predictor variables, and then adding an intercept to the model. (Multiplying two Spn variables gives zero, which does no real harm but does no good.) Try fitting such a model to these data.

**EXERCISE 11.3.** Bisgaard and Fuller (1996) give the data in Table 11.3 on the numbers of defec-tives per grille in a process examined by Chrysler Motors Engineering. The data are from a two-level fractional factorial design, see Christensen (1996, Chapter 17). The factors are A: Mold Cycle, B: Viscosity, C: Mold Temp, D: Mold Pressure, E: Weight, F: Priming, G: Thickening Process, H: Glass Type, J: Cutting Pattern. Repeat Exercise 11.2 making suitable adjustments for these data.

### 11.1.1 Poisson Regression for Rates

As illustrated in model (2), it is common to have Poisson regression models that include a predictor with a regression coefficient known to be one. These are commonly used to estimate and compare rates in different populations. Examples include comparing disease incidence rates for “exposed” versus “unexposed” groups, comparing rates of automobile accidents on stretches of different highways, and estimating the rate of fabric faults using model (1) of Example 11.1.1. The constant  $M_i$  in that model is called an *offset*, cf. Subsection 7.4.4. By including it, the parameter  $\theta$  is a *rate*; in this example it’s the number of faults per unit length of fabric. To compare disease incidence rates, the offset  $M_i$  would be the size of population  $i$ . A comparison of automobile crash rates might use the lengths of the highways as an offset.

Table 11.2: Watkins *et al.* data. The \* indicates the first observation after the third right spindle head was installed.

Table 11.3: *Grille defectives.*

Poisson regression for modeling a rate  $\theta_i$  of event occurrence relative to the window size  $M_i$  with covariate combination  $x_i$  specifies

$$y_i | \lambda_i \sim \text{Pois}(\lambda_i)$$

$$\log(\lambda_i) = \log(M_i) + x_i' \beta.$$

Here the rate is

$$\theta_i = e^{x_i' \beta},$$

because then

$$\lambda_i = M_i \theta_i.$$

Posterior inferences are available for the regression coefficients, for any rate  $e^{x' \beta}$  with covariates  $x$ , and for any *rate ratio* comparing a population with covariates  $x$  to a population with covariates  $x_*$ , i.e.,  $e^{(x-x_*)' \beta}$ .

The DIC or LPML statistics can be used to compare different regression models. In particular,

$$LPML = - \sum_{i=1}^n \log[\text{E}\{c_i(\beta)|y\}]$$

where

$$c_i(\beta) \equiv \frac{1}{f(y_i|\beta)} = y_i! \left( M_i e^{x_i' \beta} \right)^{-y_i} \exp \left( M_i e^{x_i' \beta} \right).$$

For computing, it is convenient to write this on the log scale and to use  $\theta_i$  in place of  $e^{x_i' \beta}$ ,

$$\log(c_i) = \log(y_i!) - y_i[\log(M_i) + \log(\theta_i)] + M_i \theta_i;$$

however, it is the posterior expectation of  $c_i$  that we must obtain, not the posterior expectation of  $\log(c_i)$ .

**EXAMPLE 11.1.2.** Data from a landmark study on the effects of smoking are presented in Table 11.4 (ignore the last column for the moment). The participants were male British doctors surveyed in 1951 to collect data on age and whether they smoked tobacco. Ten years later the number of deaths from coronary heart disease (CHD) and the number of person-years on study in each group were determined. We want to compare death rates for smokers and nonsmokers, controlling for age. Figures 11.1 and 11.2 give empirical death rates by age on the original and log scales, respectively.

Table 11.4: *Ten-year mortalities from coronary heart disease and posterior rate ratios.*

Age	Smokers		Nonsmokers		Post. rate ratio med.(95% PI)
	Deaths	Person-years	Deaths	Person-years	
35-44	32	52407	2	18790	4.0 (1.7, 11.0)
45-54	104	43248	12	10673	2.4 (1.6, 3.7)
55-64	206	28612	28	5710	1.6 (1.2, 2.2)
65-74	186	12663	28	2585	1.2 (0.9, 1.5)
75-84	102	5317	31	1462	0.98 (0.7, 1.5)

From Figure 11.1, except in the 75 to 84 age range, the empirical mortality rates from coronary heart disease were higher in smokers than nonsmokers and the differences are increasing with age. Somewhat ironically, Figure 11.2 indicates that the empirical rate ratios are decreasing with age. For young people, the relative risk of smoking is greater but the actual risk is very small. As age increases, the relative risks decrease, while the actual risks increase as do the risk differences. From Figure 11.2, parabolas for log mortality look like they would fit the data well.

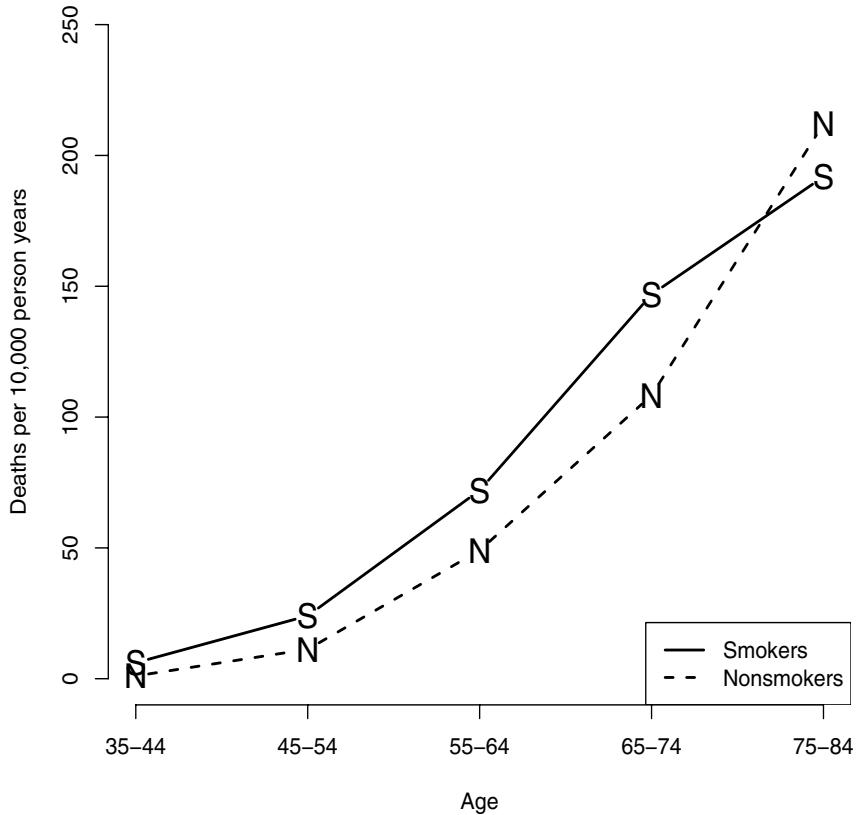


Figure 11.1: Empirical death rates from CHD for smokers and nonsmokers.

Let  $\theta_i$  denote the mortality rate in group  $i$ ,  $i = 1, \dots, 10$ , where the first five groups are smokers and the last five are nonsmokers. We fit a Poisson regression model with

$$\log(\theta_i) = \beta_1 + \beta_2 A_i + \beta_3 S_i,$$

where  $A_i$  is the age (enumerated as 1–5 for the five age groups) and  $S_i$  is the smoking status (1 for smoke, 0 for no smoke) of group  $i$ . Independent  $N(0, 100)$  priors were used for regression coefficients. Five chains were run with different starting values and we kept every 50th iterate to reduce autocorrelation. Posterior approximations were based on the last 100,000 retained iterates from each chain after the first 10,000 iterates were discarded as burn-in. The DIC is 130.3. In comparison, a model with separate log-linear trends for smokers and nonsmokers,  $\log(\theta_i) = \beta_1 + \beta_2 A_i + \beta_3 S_i + \beta_4 S_i A_i$ , has DIC = 123.0. A model with separate quadratic trends, where  $\log(\theta_i) = \beta_1 + \beta_2 A_i + \beta_3 A_i^2 + \beta_4 S_i + \beta_5 S_i A_i + \beta_6 S_i A_i^2$ , has a much smaller DIC value of 68.2. The last column of Table 11.4 presents posterior medians and 95% PIs for rate ratios comparing smokers to nonsmokers under the quadratic trend model, i.e.,  $\theta_j / \theta_{j+5}$ ,  $j = 1, \dots, 5$ . The median risk ratios decrease with age, as do the widths of the 95% intervals.

Table 11.5 gives posterior percentiles of the mortality rates  $\theta_j$  and for the rate differences  $\theta_j - \theta_{j+5}$ . Clearly, fitted risks increase with age. There is almost no overlap between 95% intervals as ages increase for either smokers or nonsmokers. It is often thought that the key contribution of formal statistical analysis is the evaluation of variability. Variability (width of the intervals) increases

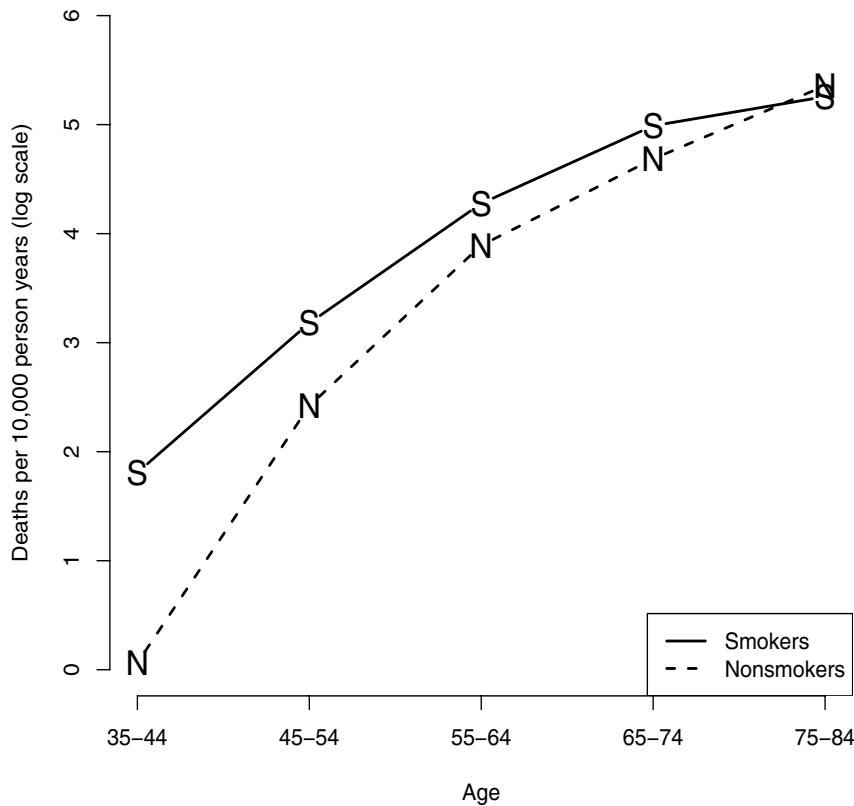


Figure 11.2: Empirical death rates from CHD for smokers and nonsmokers. Log scale.

markedly with age. We observed from Figure 11.1 that the empirical risk differences are increasing up to the last age group. From Table 11.5 we find that the median risk differences only increase in the first three age groups. However, we also see a large increase in variability at the fourth age group to the point where we are not completely sure that smokers show any increased risk for this age group. For the oldest group, there is another jump in variability to the point that the negative estimate of risk difference has no importance.

**EXERCISE 11.4.** Fit the three models discussed above for the coronary heart disease data and from each, plot posterior mean estimates of the death rate per 10,000 person years with the empirical rates. Discuss the results. Adapt the following WinBUGS code.

```
model{
  for(i in 1:10){
    y[i] ~ dpois(lambda[i])
    lambda[i] <- M[i]*theta[i]
    log(theta[i]) <- beta[1]+beta[2]*A[i]+beta[3]*A[i]*A[i]
      +beta[4]*S[i]+beta[5]*A[i]*S[i]+beta[6]*A[i]*A[i]*S[i]
  }
  for(i in 1:6){beta[i] ~ dnorm(0,0.01)}
  for(i in 1:5){RateRatio[i] <- theta[i] / theta[i+5]}
```

Table 11.5: Ten-year posterior mortality rates and rate differences, quadratic model.

Age	Smoker	Par.	2.50%	50%	97.50%
35-44	Yes	$\theta_1$	4.259	5.752	7.604
45-54	Yes	$\theta_2$	21.69	24.69	27.94
55-64	Yes	$\theta_3$	64.92	72.19	80.04
65-74	Yes	$\theta_4$	130.1	143.7	158.2
75-84	Yes	$\theta_5$	162.2	194.7	231.6
35-44	No	$\theta_6$	0.558	1.452	3.294
45-54	No	$\theta_7$	6.749	10.31	14.84
55-64	No	$\theta_8$	33.54	45.07	59.2
65-74	No	$\theta_9$	93.86	121.2	153.9
75-84	No	$\theta_{10}$	138.9	200.2	278.2
		$\theta_1 - \theta_6$	1.976	4.24	6.362
		$\theta_2 - \theta_7$	8.967	14.36	19.21
		$\theta_3 - \theta_8$	11.25	27.1	41.11
		$\theta_4 - \theta_9$	-12.81	22.49	53.49
		$\theta_5 - \theta_{10}$	-89.78	-5.494	66.23

```
}
list(A=c(1,2,3,4,5,1,2,3,4,5),
S=c(1,1,1,1,1,0,0,0,0,0),
y=c(32,104,206,186,102, 2,12,28,28,31),
M=c(52407,43248,28612,12663,5317,18790,10673,5710,2585,1462))
```

EXERCISE 11.5. For the coronary heart disease data, treat age as a categorical variable. (a) Compare results of a main effects model with those obtained for the quadratic trend model reported in Table 11.5. (b) Calculate the pseudo Bayes factor based on LPMLs to compare the main effects model with a smoking-age interaction model. Also use DIC to compare these models. (c) Discuss your findings. Which model is preferred among all those considered?

## 11.2 Over-Dispersion and Mixtures of Poissons

Sometimes count data display more variability than is appropriate for Poisson data. This is known as *over-dispersion*. If we have a random sample of count data  $y_1, \dots, y_n$ , there is an easy way to check for over-dispersion. Recalling that the mean and variance of a Poisson have the same value, simply look at  $s_y^2/\bar{y}_.$ , the ratio of the sample variance to the sample mean. If this is substantially greater than 1, there is evidence of over-dispersion.

Next we illustrate how over-dispersion might occur. Our discussion of over-dispersion differs from that in typical discussions of generalized linear models. Those typically involve adding another parameter to the Poisson's discrete density. We motivate over-dispersion using mixture models, i.e., randomly choosing the parameter of the distribution. Mixture models are discussed more fully in Subsection 15.1.1.

We once needed to send a silicon chip to a lab to detect the number of faults  $y$  on the chip. We assigned this task to our deadbeat brother-in-law who found two labs known to have different rates,  $\theta_1$  and  $\theta_2$ , of finding faults. Being the good decision maker that he is, he simply flipped a coin to decide on a lab. Before skipping town for nefarious reasons too embarrassing to mention, he left a note telling us only the value of  $y$ .

In our ignorance, we might define  $E(y) \equiv \theta$  and we might even tentatively assume that  $y \sim \text{Pois}(\theta)$  implying that both the mean and variance are  $\theta$ . We now demonstrate that  $y$  has over-dispersion, thus invalidating the Poisson model.

If we knew all of the details of our brother-in-law's activities, we could make a better model. If we knew that the chip was sent to lab  $k$ , a reasonable model for the number of faults as measured by lab  $k$  might be  $W_k \sim \text{Pois}(\theta_k)$ . Let  $\xi$  be the outcome of flipping the coin, so  $\xi \sim \text{Bern}(0.5)$ . Not knowing the lab, a better model for our actual observation is the *mixture model*

$$y \sim \xi W_1 + (1 - \xi) W_2. \quad (1)$$

It follows that our parameter of interest has  $E(y) \equiv \theta = 0.5\theta_1 + 0.5\theta_2$ . Using Proposition B.2 and conditioning on  $\xi$ , it is not too difficult to show that  $\text{Var}(y) = (0.5\theta_1 + 0.5\theta_2) + 0.25(\theta_1 - \theta_2)^2 > \theta$ , so the observation is subject to over-dispersion. Another way to write the mixture model is

$$\begin{aligned} y|\theta &\sim \text{Pois}(\theta), \\ \theta &= \xi\theta_1 + (1 - \xi)\theta_2, \\ \xi &\sim \text{Bern}(0.5). \end{aligned} \quad (2)$$

This form makes it clear that the Poisson parameter is random. In either model (1) or (2) the density for  $y$  is

$$f(y|\theta_1, \theta_2, 0.5) = 0.5f(y|\theta_1) + 0.5f(y|\theta_2)$$

where  $f(y|\theta)$  is the Poisson density.

The real point is that we do not know how this  $y$  value came to us. We demonstrated that over-dispersion can be modeled by taking the Poisson parameter value  $\theta$  as a random variable. Our next example illustrates our basic model for Poisson regression with over-dispersion, which is

$$\begin{aligned} y_i|\lambda_i &\stackrel{\text{ind}}{\sim} \text{Pois}(\lambda_i), \quad i = 1, \dots, n \\ \log(\lambda_i) &= x_i'\beta + z_i'\eta, \end{aligned}$$

where  $\eta$  is a random vector, typically assumed to be multivariate normal. If  $x_i'\beta$  contains an intercept,  $\eta$  should be centered at 0. Often  $z_i$  is used to indicate "group membership," so each group gets a separate  $\eta_k$  term.

**EXAMPLE 11.2.1.** We return to the Ache armadillo hunting of Section 1.5. The data consist of the daily armadillo kills of 38 Ache men over several treks. The 1,302 total observations  $y_{ij}$  were indexed  $i = 1, \dots, 38$ ,  $j = 1, \dots, n_i$ . Interest focused on how age affects daily kill success. We assume the log kill rate is a quadratic function of  $a_i$ , age in years, plus a subject-specific, normally distributed random effect  $\eta_i$ . The random effect accounts for correlation of an individual's daily kills over the several hunting trips and can be viewed as the innate ability of the hunter.

The sampling model is

$$\begin{aligned} y_{ij}|\lambda_i &\stackrel{\text{ind}}{\sim} \text{Pois}(\lambda_i), \quad i = 1, \dots, 38; j = 1, \dots, n_i, \\ \log(\lambda_i) &= \beta_1 + \beta_2(a_i - \bar{a}) + \beta_3(a_i - \bar{a})^2 + \eta_i \\ \eta_i|\tau &\stackrel{\text{iid}}{\sim} N(0, 1/\tau), \quad i = 1, \dots, 38. \end{aligned}$$

Here  $\lambda_i$  is the mean daily kill rate for individual  $i$  and  $\bar{a}$  is the average age of the 38 hunters. The ages  $a_i$  are fixed, known constants.

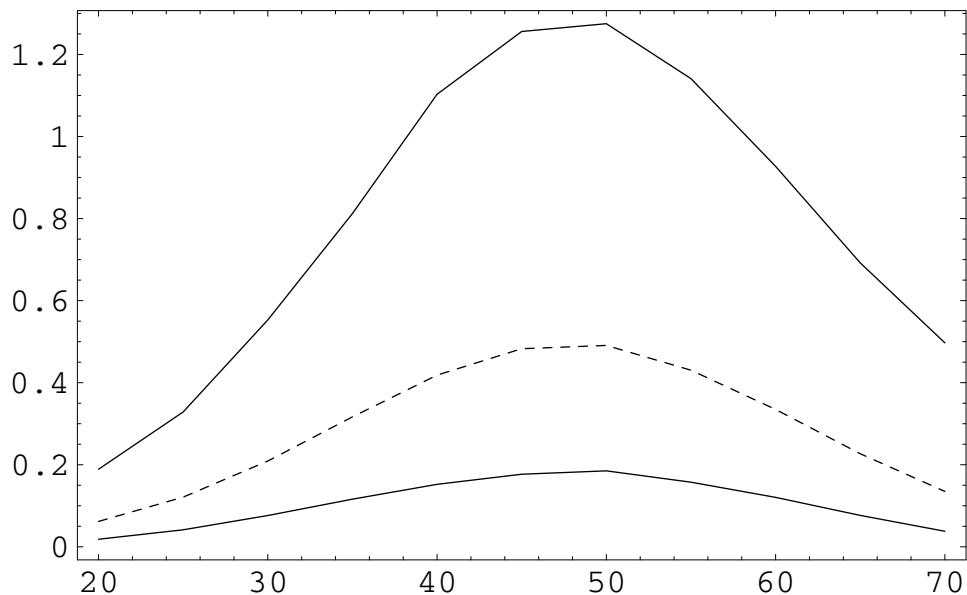
The parameters  $\beta_1$ ,  $\beta_2$ ,  $\beta_3$ , and  $\tau$  are given independent reference prior distributions:  $\beta_i \sim N(0, 1000)$ ,  $i = 1, 2, 3$ , and  $\tau \sim \text{Gamma}(0.001, 0.001)$ .

In addition to information on the regression model parameters, we wish to obtain pointwise prediction intervals for the unobservable kill rates of men of various ages not included in the data.

Summary information on parameters is presented in Table 11.6. The quadratic coefficient median is  $\tilde{\beta}_3 = -0.0027$  and from the probability interval it is clearly nonzero. A simple (but

Table 11.6: *Armadillo kills: posterior medians and probability intervals.*

Parameter	Median	95% PI
$\beta_1$	-0.7147	(-1.007, -0.433)
$\beta_2$	0.01368	(-0.002525, 0.03085)
$\beta_3$	-0.002683	(-0.004007, -0.001459)
$\sigma$	0.4252	(0.2731, 0.658)

Figure 11.3: *Estimated mean daily kill by age with 95% PI.*

Bayesianly unjustified) method of examining the relationship between kill rate and age is to exponentiate  $\log(\lambda)$  after plugging in estimates of the regression coefficients:

$$\hat{\lambda}(a) = \exp\{-0.7147 + 0.01368(a - \bar{a}) - 0.002683(a - \bar{a})^2\}.$$

This estimate is not necessarily the posterior median, mean, or mode.

Posterior medians for the mean daily kill  $\lambda(a)$  and probability intervals were computed every five years from ages 20 to 70 to obtain Figure 11.3. The range of ages in the actual data is 20 to 66 years. We see that the average kill rate increases with age up until about 50, perhaps reflecting that hunting experience increases the chance of killing an armadillo, but then declines as the hunter enters his “golden years.” When the model was refit using independent (infinite and improper) uniform priors on all model parameters the resulting posterior inferences were almost identical to those presented above.

**EXERCISE 11.6.** To obtain Figure 11.3, modify the following WinBUGS code that fits a *linear* association for age. The data on our website are written with 1,302 cases, one for each pair  $ij$  in the model. The variable  $id$  identifies a hunter, thus giving the index  $i$  in the model. Note the distinct use of “age” and “ages.”

```
model{
  for(k in 1:1302){
    kills[k] ~ dpois(lambda[k])
```

```

log(lambda[k]) <- beta1+beta2*(age[k]-mean(age[]))+delta[id[k]]
}
for(i in 1:38){ delta[i] ~ dnorm(0,tau) }
for(i in 1:11){
  d[i] ~ dnorm(0,tau)
  log(r[i]) <- beta1 + beta2*(ages[i]-mean(age[])) + d[i]
}
beta1 ~ dnorm(0,0.0001)
beta2 ~ dnorm(0,0.0001)
tau ~ dgamma(0.0001,0.0001)
}
list(ages=c(20,25,30,35,40,45,50,55,60,65,70))

```

EXAMPLE 11.2.2. Example 11.1.1 examined fabric faults. An over-dispersion version of model (11.1.2) has the sampling model  $y_i|\lambda_i \stackrel{iid}{\sim} \text{Pois}(\lambda_i)$ ,  $i = 1, \dots, 32$ ,

$$\log(\lambda_i) = \beta_1 + \log(M_i) + \eta_i, \quad \eta_i | \tau \stackrel{iid}{\sim} N(0, 1/\tau).$$

We used priors

$$\beta_1 \sim N(0, 1000) \quad \perp \!\!\! \perp \quad \tau \sim \text{Gamma}(0.001, 0.001).$$

Our primary interest is in whether we need the over-dispersion model. If  $\tau = \infty$  there is no over-dispersion. In fact, the posterior mean of  $\tau$  is 10.95 with posterior median of 8.69. The 95% PI is (3.46, 31.68). These are all substantial numbers but are they large enough to indicate that over-dispersion is not a problem? Based on the deviance information criterion, it seems we need the model with over-dispersion. The DIC for this model is 171.7, whereas the DIC for the same model without over-dispersion, model (11.1.2), is 189.8. Is a difference of 18 a substantial improvement? Well, the DIC for model (11.1.3) is 191.8, slightly worse than for model (11.1.2). Although there is some reason to believe that DIC underpenalizes random effects models like this one, the decrease from 190 to 172 may be large enough to be meaningful.

EXERCISE 11.7. Completely redo Example 11.1.1 incorporating a random effect for over-dispersion. Begin with the following WinBUGS code.

```

model{
  for(i in 1:32){
    y[i] ~ dpois(lambda[i])
    log(lambda[i]) <- beta1 + log(M[i]) + delta[i]
    delta[i] ~ dnorm(0,tau)
  }
  beta1 ~ dnorm(0,0.001)
  tau ~ dgamma(0.001, 0.001)
  junk <- ID[1]
}

```

EXERCISE 11.8. Redo Exercise 11.2 incorporating a random effect for over-dispersion. Note that the random effect should depend only on the run.

EXERCISE 11.9. Redo Exercise 11.3 incorporating a random effect for over-dispersion.

### 11.2.1 Zero-Inflated Poisson Data

Count data may be over-dispersed when they exhibit more zeros than expected under a Poisson sampling model. The zero-inflated Poisson (ZIP) distribution accommodates excess zeros through the two-component mixture model

$$y \sim \xi W,$$

where  $\xi$  is a Bernoulli random variable with unknown mixing proportion  $\pi$  and  $W$  is an independent  $\text{Pois}(\theta)$ . The density of the ZIP distribution is

$$f(y|\pi, \theta) = \begin{cases} (1-\pi) + \pi e^{-\theta}, & y=0 \\ \pi e^{-\theta} \theta^y / y!, & y=1, 2, \dots \end{cases}$$

The mean and variance are:

$$\begin{aligned} E(y|\pi, \theta) &= \pi\theta \\ \text{Var}(y|\pi, \theta) &= \pi\theta[1 + (1-\pi)\theta]. \end{aligned}$$

Compared to the ordinary  $\text{Pois}(\theta)$ , the mean of the ZIP model is smaller (since the excess zeros pull the mean down), but the variance is larger than the mean, so it provides a potential model for over-dispersed count data.

Noting that a  $\text{Pois}(0)$  puts probability one on seeing zero, a sample of zero-inflated Poisson regression specifies

$$\begin{aligned} y_i | \lambda_i &\stackrel{\text{ind}}{\sim} \text{Pois}(\lambda_i) \\ \lambda_i &= \xi_i \theta_i \\ \xi_i | \pi &\stackrel{\text{iid}}{\sim} \text{Ber}(\pi) \\ \log(\theta_i) &= x_i' \beta. \end{aligned}$$

A beta prior can be placed on  $\pi$ , with independent normal priors for the regression coefficients.

**EXAMPLE 11.2.3. Foot-and-Mouth Disease.** Branscum et al. (2008) analyzed data on reported cases of foot-and-mouth disease (FMD) in each province of Turkey over the eight years from 1996 to 2003. We consider data from 1998 with the goal of assessing the difference in FMD incidence between the eastern and western regions of the country. The histogram in Figure 11.4 shows that the majority of provinces (44 out of 66) had no reported FMD cases in 1998.

We consider both an ordinary Poisson regression and a ZIP regression using region and size of cattle population for each province as covariates. The cattle populations were standardized by subtracting the mean and dividing by the standard deviation. Independent  $N(0, 1000)$  priors were used for the regression coefficients.

WinBUGS produced an error message for the ordinary Poisson regression data analysis. We could fiddle with WinBUGS to get it to run. Instead, we take this opportunity to illustrate the use of `proc mcmc` in the *SAS* statistical software package, a procedure that was first made available in version 9.2 of SAS.

### 11.2.2 SAS Analysis of Foot-and-Mouth Disease Data

The data include the 1998 disease counts for each province `FMD1998`, an indicator variable `EasternTurkey`, and the standardized cattle population per province `stdcattle`.

The primary snippet of SAS code is

```
proc mcmc data=FMDdata nbi=15000 nmc=1000000 dic propcov=quanew;
parms beta1 0 beta2 0 beta3 0;
```

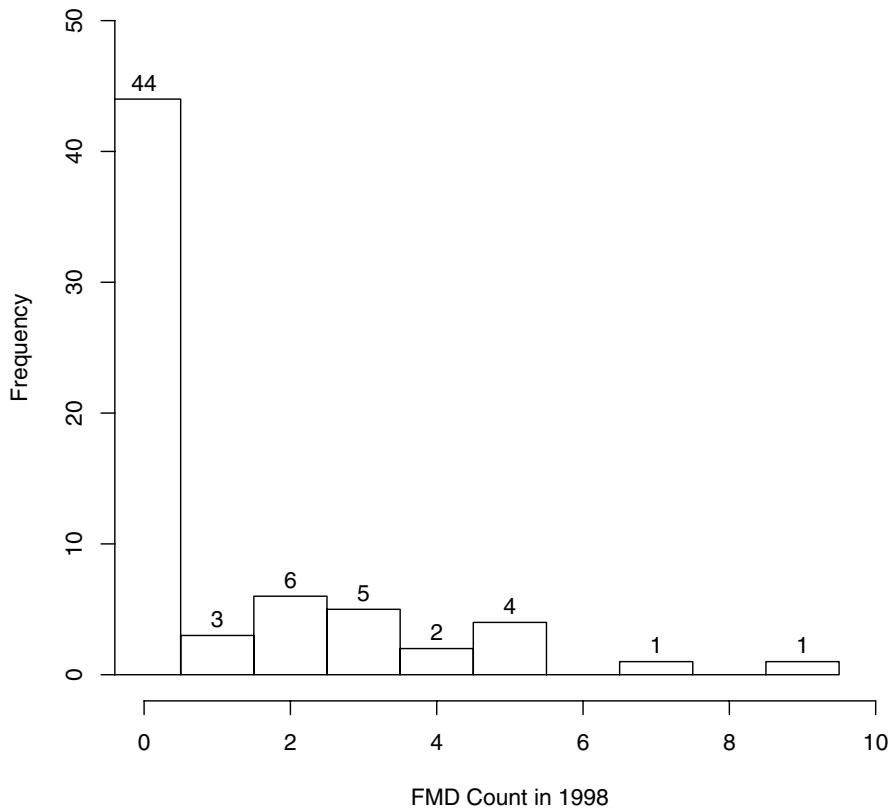


Figure 11.4: Histogram of FMD in 66 provinces of Turkey in 1998.

```

prior beta: ~ normal(0,var=1000);
lambda = exp(beta1 + beta2*stdcattle + beta3*EasternTurkey);
model FMD1998 ~ poisson(lambda);
run;

```

The first line identifies the data, FMDdata, specifies a burn-in of 15,000 (nbi) and an MCMC sample size of 1,000,000 (nmc). It also specifies that we want the *DIC* and propcov specifies the initial covariance matrix for the normal proposal distribution in Metropolis-Hastings sampling. The second line identifies the parameters and specifies initial values for them. The third line gives the prior on the parameters. The fourth line specifies the log-linear model and the fifth line specifies that the 1998 counts are to have a Poisson distribution with mean determined by the log-linear model.

For the ZIP regression the key changes are in the last two lines. Rather than specifying the model as having a standard sampling distribution like the Poisson, we specify the log-likelihood function and then identify the model as one with a general log-likelihood.

```

proc mcmc data=FMDdata nbi=15000 nmc=1000000 dic propcov=quanew;
parms beta1 0 beta2 0 beta3 0 pi .5;
prior beta: ~ normal(0,var=1000);
prior pi ~ uniform(0,1);
theta = exp(beta1 + beta2*stdcattle + beta3*EasternTurkey);

```

```
llike=log((1-pi)*(FMD1998 eq 0) + pi*pdf("poisson",FMD1998,theta));
model general(llike);
```

The DIC value for the ordinary Poisson regression is 231, but the ZIP model is preferred based on its much smaller DIC of 172. A portion of the SAS output is displayed below. Unlike WinBUGS, SAS provides highest posterior density (HPD) intervals, cf. Section 2.4.

Parameter	mean	sd	95% PI		95% PI	
			Equal-Tail		HPD	
beta1	1.01	0.17	0.66	1.32	0.68	1.34
beta2	0.19	0.11	-0.02	0.42	-0.02	0.41
beta3	0.05	0.37	-0.70	0.74	-0.67	0.76
pi	0.63	0.06	0.50	0.75	0.50	0.75

The interval estimates of  $\beta_3$  are close to being symmetric about 0, and the posterior mean is 0.05. The data do not indicate a meaningful difference in mean FMD occurrence for eastern versus western Turkey in 1998.

### 11.3 Longitudinal Data

The data in Table 11.2 are described in Exercise 11.2. There are three observations at every set of experimental conditions. It is implicit from the information on the Heads that the observations were taken at consecutive times, say  $t_1 = 1$ ,  $t_2 = 2$ , and  $t_3 = 3$ . We can easily incorporate time trends into the model of Exercise 11.2. For a simple linear time trend

$$y_{ij} | \lambda_{ij} \stackrel{\text{ind}}{\sim} \text{Pois}(\lambda_{ij}), \quad i = 1, \dots, 19, \quad j = 1, 2, 3,$$

$$\begin{aligned} \log(\lambda_{ij}) &= \beta_1 \text{Ch}_i + \beta_2 \text{Cl}_i + \beta_3 \text{Spd}_i + \beta_4 \text{P}_i + \beta_5 \text{RA}_i \\ &\quad + \beta_6 \text{Spn}_{1i} + \beta_7 \text{Spn}_{2i} + \beta_8 \text{Spn}_{3i} + \beta_9 \text{Spn}_{4i} + \delta_i + \gamma_1 t_j, \\ \delta_i | \tau &\stackrel{\text{iid}}{\sim} N(0, \tau). \end{aligned}$$

The random effect  $\delta_i$  not only models over-dispersion but also, given the regression parameters, causes a correlation among the elements of  $y_i \equiv (y_{i1}, y_{i2}, y_{i3})'$ . Note that the  $y_i$  vectors are independent given the regression parameters. The time trend  $\gamma_1 t_j$  did not include an intercept because the model already includes the equivalent of an intercept with the four  $\text{Spn}_{ki}$  indicator variables. Adding a quadratic term  $\gamma_2 t_j^2$  to the model would allow for a curved time trend.

EXERCISE 11.10. Fit the longitudinal model given above. Generalize the model to allow a different slope for each run. With a single slope for all runs, add a single quadratic term for all runs. Also, implicit from the information on Spindles, we know the sequential order of all of the observations. Eliminate the within runs time trends but add an overall time trend.

---

## Chapter 12

---

# Time to Event Data

---

In this chapter we consider the analysis of time to event data. Examples include (i) the time until death after diagnosis with leukemia, (ii) the time it takes to get sick after infection with a virus, e.g., the human immunodeficiency virus (HIV), (iii) the time until a machine breaks down after being installed, and (iv) the time it takes to learn a new skill after the beginning of instruction. *Survival analysis* is the term used to describe the analysis of time to event data in biological contexts. *Reliability analysis* is often used for non-biological applications. A majority of our authors do survival analysis, so we often slip into that terminology. Only the first of these four examples actually deals with survival. The second is clearly biological. The third is about reliability. The fourth could swing either way.

In Section 1 we discuss the basic vocabulary of survival analysis including definitions of hazard and survival functions, and censoring. Sections 2 and 3 discuss one- and two-sample problems from exponential, Weibull, and log-normal populations. Section 4 addresses issues of plotting survival, density, and hazard functions in R.

### 12.1 Introduction

As of 2004 the two most highly cited research articles in Statistics were both on survival analysis, see Ryan and Woodall (2005). The most common goal of survival analysis is to compare survival prospects among different populations. That sounds suspiciously like comparing two or more populations, a task that we have repeatedly addressed. What could justify an entire chapter on time to event data? Two things. First is a technical issue. Time to event data are often only partially observed. We often know that a unit (a person, a car, a refrigerator) was operative (alive, healthy, working) up to a certain time but do not know exactly when it failed or would fail. That is called *censoring*. Censoring complicates all of the technical issues involved in analyzing the data. Second is that time to event data are considered extremely important by society. A large proportion of medical studies collect time to event data. There is broad interest in assessing time of survival or time to cure for people diagnosed with adverse medical conditions. Human medicine is a huge area of research and survival analysis is an integral part of such research (as is logistic regression and associated methods).

Time to event data are distinguished by two features: (i) they are positive and (ii) they are often censored. What's the big deal about data being positive? In Chapters 5, 9, and 10 we used normal distributions to analyze measurement data and most measurement data are positive. Time to event data are often skewed, so we would need to take, say, a log transformation before analyzing them as normal. But unlike the Diasorin scores of Chapters 5 and 9, people really want the analysis on the original measurement scale. People want to know the actual time of survival, not the log of the time of survival. Let  $T$  denote the time to the event. Our favorite model for time to event data is  $\log(T) \sim N(\mu, 1/\tau)$  but now with an emphasis on interpretations related to  $T$ . Alternatively, the Exponential distribution has nice theoretical properties for time to event analysis. We could also model  $T$  as Weibull or Gamma, both being generalizations of the Exponential. These four are the fundamental parametric distributions used for event time data.

Time to event data often depend on covariates. Risk factor information such as age at the beginning of the study or treatment regimen are often associated with survival outcomes. In the next chapter we consider *accelerated failure time (AFT)* models that incorporate both covariates and an array of possible survival distributions like the log-normal, the Exponential, and the Weibull. These models allow us to use a common notation both for the models themselves and for the inferences we make. They also make model comparison and prior selection easier. In addition, we consider the *proportional hazards (PH)* model, a highly cited semi-parametric model that provides greater flexibility than standard parametric models.

Censoring creates a new technical issue for obtaining the likelihood function. Studies are only funded for a finite time. When funding stops, data collection stops. Generally, there will be individuals who have not yet “died,” “learned the skill,” or “become sick.” Moreover, we cannot keep people in cages to observe them. People drop out of studies for a variety of reasons: they move away, they dislike the person they are dealing with, they die from a competing risk like a car crash, or they die from the disease but the study does not find out. Typically we have exact times for some events, but for others all we will know is that the event occurred sometime after they were last seen. These are examples of censoring. The contribution of a censored individual to the likelihood function is different than for one whose event time is known.

One attractive thing about time to event analysis is that it focuses on observables. For example, with two competing cancer treatments, a time to event analysis is much more likely to care about the relative probabilities of surviving past 5 years than to worry about the means of the two survival distributions. The parameters of interest tend to be probabilities of various events (e.g., survival past 5 years) rather than means and variances.

### 12.1.1 Survival and Hazard Functions

With time to event data, the primary object of analysis is the *survival function*  $S(t)$ , defined as

$$S(t) \equiv \Pr(T > t).$$

The survival function is just 1 minus the cumulative distribution function, i.e.,  $S(t) = 1 - F(t)$ . The survival function is also called the *reliability function* and denoted  $R(t)$ . We use the survival terminology and notation throughout. Time to event data must always be nonnegative and they are usually, though not always, regarded as continuous. The survival function (also called the *survival curve*) is a non-increasing function. At the time origin,  $S(0) = 1$  (everybody is alive), and as  $t$  gets large,  $S(t)$  tends to 0 (everything/everybody eventually breaks down).

The hazard function at a specific time  $t$  is the instantaneous rate of event occurrence among the population that is still at risk at time  $t$ . Technically, if a waiting time  $T$  has as density function  $f(t)$ , the hazard function is

$$\begin{aligned} h(t) &= \lim_{\Delta \rightarrow 0} \frac{\Pr(t < T \leq t + \Delta | T > t)}{\Delta} \\ &= \lim_{\Delta \rightarrow 0} \frac{\Pr(T \in (t, t + \Delta] \cap T > t) / \Pr(T > t)}{\Delta} \\ &= \lim_{\Delta \rightarrow 0} \frac{\Pr(T \in (t, t + \Delta])}{\Delta} / \Pr(T > t) \\ &= \lim_{\Delta \rightarrow 0} \frac{\int_t^{t+\Delta} f(u) du}{\Delta} / \Pr(T > t) \\ &= \frac{f(t)}{S(t)}. \end{aligned} \tag{1}$$

The third equality follows because  $T \in (t, t + \Delta]$  implies  $T > t$  and the last equality follows from the Fundamental Theorem of Calculus and the definition of a derivative. If  $\Pr[T \leq T_0] = 1$ , the hazard

function is undefined for  $t \geq T_0$ . The most direct way to think of the hazard is as

$$h(t) \doteq \Pr(T \in (t, t + \Delta] \mid T > t),$$

the probability someone will have their event in the next small window of time given that they are alive and well at time  $t$ . Knowing any one of the density  $f$ , cdf  $F$ , survival function  $S$ , or hazard function  $h$  determines the other three. The “*incidence density*” in Epidemiology is a simple estimate of the hazard function.

The most commonly applicable hazard function for the entire life of an object is a bathtub shape. When the unit is put in service (you are born), there is a good chance of early failure (death). As the kinks are worked out, or perhaps more correctly, as it becomes clear that certain faults are not present (you escape catastrophic birth defects and childhood diseases), the hazard of failure decreases with age. The hazard is flat and low during the mature operating life of the object. Finally, with increased age, the hazard of breakdown increases as the unit wears out. We will see that Exponential distributions have a flat hazard function, so they are particularly useful for data that look only at the mature operating life of objects.

In time to event data analysis, the curves  $S(t)$  and  $h(t)$  are objects of primary interest, as opposed to some vector of parameters. Focusing on survival functions is essentially focusing on prediction. We also focus attention on median survival times rather than on mean survival times. That is because survival distributions are typically skewed rather than symmetric, making the median a more stable basis for interpretation than the mean.

Survival and hazard functions often depend on covariates. Our models in Chapter 13 with covariates have corresponding regression coefficients that we will estimate and interpret. But regression coefficients will be secondary parameters as they are less directly interpretable than hazards, survival probabilities, and median survival times.

### 12.1.2 Censoring

Censoring is the key technical innovation in this chapter and the next. Let  $C$  be a random variable that denotes the time at which a censoring mechanism kicks in. What we actually observe in time to event studies is either the event time  $T$  or the censoring time  $C$ , whichever is smaller. The observed data are

$$y = \min(T, C).$$

In addition, we usually get information on whether  $y$  is an actual event time or a censored observation. Define an indicator random variable for noncensoring

$$\delta = \begin{cases} 1 & T \leq C \\ 0 & T > C \end{cases},$$

so  $\delta$  is 1 if we observe an actual event time and 0 if we observe when an observation is censored. (Somewhat bafflingly,  $\delta$  is often called a censoring indicator even though it indicates noncensoring.) There is information about  $T$  even in censored observations. If  $(y, \delta) = (y_0, 0)$ , we know that  $T > y_0$ .

To simplify our lives, we assume that  $T$  and  $C$  are independent. For example, if older people are both more likely to die and to be censored, we lack independence (unless the model accounts for age). If people who are near death get pulled from the study, time of death  $T$  and time of censoring  $C$  are clearly correlated. More generally, if the condition of the patient can cause them to be censored,  $T$  and  $C$  will not be independent (unless one can make the difficult argument that the condition that caused censoring was not related to the disease in question or any other aspect of the study). Decisions about medical treatments involve obvious ethical considerations that are beyond the scope of our discussion. However, our likelihood function depends crucially on the assumption that  $T$  and  $C$  are independent.

Another assumption is that the censoring distribution, say  $G(c) = \Pr(C \leq c)$ , does not depend on any of the same parameters as  $S(t)$ . This is called *noninformative censoring* or *uninformative*

*censoring.* For example, suppose we measure survival in days and see one uncensored observation  $y = 25$ . Then 25 would be our best estimate of the median time of survival. However, if we observe one censored observation  $(y, \delta) = (25, 0)$ , the fact that the observation is censored would typically make our best estimate of the median something larger than 25. Now suppose we have informative censoring so that the time to event and censoring distributions have a parameter in common. Specifically, let  $T \sim \text{Pois}(\theta)$  and  $C \sim \text{Pois}(\theta + 5)$ . Seeing a censored observation of 25 now makes us think that the median time to event should be about 20 because of the informative censoring. Non-informative censoring is assumed in most survival analysis studies. (The key to this example is that the censoring distribution is the event distribution shifted to the right. Although a Poisson might be used to model, say, the number of days until an event occurs [when a geometric distribution does not fit], it was used here primarily because it is easy to shift a Poisson to the right.)

Censoring is a bad thing! Clearly, it would be more informative to observe every unit until the event of interest actually occurs. But often the objects of study have sufficiently high survivability that in the normal course of events, many survive past the time we have to study them. Most people live a long time, so do most television sets. As a result, we often accelerate the conditions under test to levels not normally experienced by the objects of study. If we want to test whether aspartame causes cancer in rats, we do not give the rats diet soda to drink in amounts that a rat would normally drink—we force into their systems amounts of aspartame that would be equivalent to a human drinking cases of soda every day. Similarly, we might test a television by running unusually high voltages through it. The idea is to induce enough stress so that it causes a reasonable number of events to occur in the time allotted for the study. The down, side is that one needs to extrapolate the results obtained at the high stress conditions down to applications at normal stress conditions. Extrapolation is always tricky since it is difficult to predict behavior at places where you have not collected data.

Our definition of censoring is the most common form of censoring, also known as *right censoring*. Sometimes observations are partially lost because they are too small to be seen. For convenience, many medical studies begin at the time of diagnosis. For infectious diseases, it would be more appropriate to begin them at the time of infection. However, the time of infection is typically only known to be something earlier than the time of diagnosis. The time of infection is then *left censored*. Occasionally, event times are only known to occur inside some time interval, say, between visits to a medical clinic. Such data are *interval censored*. In reality, all measurement data are interval censored. All measurements, including time measurements, are numbers in a small interval determined by the accuracy of the measuring instrument. As long as the intervals are narrow, we typically ignore the interval censoring.

### 12.1.3 The Likelihood

In practice, the survival distribution is unknown and we need to estimate it from data. We can either assume that the survival distribution belongs to a parametric family, e.g., log-normal, Exponential, Weibull, or Gamma, or we can take a nonparametric approach to estimating the survival curve. In either case, we assume independent observations  $(y_i, \delta_i)$  on  $i = 1, \dots, n$  units. Eventually, we also assume the availability of covariate information (predictor variables) on each unit, vectors  $x_i$ ,  $i = 1, \dots, n$ . The covariates can be things like the age of the unit at the beginning of the study, indicators for a treatment given to the unit, etc. We continue to write  $y = (y_1, \dots, y_n)'$  and now write the collection of non-censoring indicators as  $\delta = (\delta_1, \dots, \delta_n)'$ . The complete data are  $D \equiv (y, \delta)$ .

Being traditional statisticians, we introduce parameters so we have something to estimate. For each event time  $T_i$ , assume that the density  $f_i(t|\theta)$  is known except for the parameter vector  $\theta$ . The survival and hazard functions are also members of parametric families,  $S_i(t|\theta)$  and  $h_i(t|\theta)$ , respectively. If the data are a censored random sample from a single population, just remove the subscript  $i$  from the three functions. If individual  $i$  has covariate/risk-factor information  $x_i$ , typically  $f_i(t|\theta) \equiv f(t|x_i, \theta)$ ,  $S_i(t|\theta) \equiv S(t|x_i, \theta)$ , and  $h_i(t|\theta) \equiv h(t|x_i, \theta)$ .

The likelihood function is the product of terms for observed survival times and terms for observed censored times. Specifically, it is the product of the densities for all of the actual observed survival times multiplied by the product of the probabilities for all of the censored observations. The likelihood function based on all the data is

$$L(\theta | D) \propto \prod_{i=1}^n [f_i(y_i | \theta)]^{\delta_i} [S_i(y_i | \theta)]^{1-\delta_i}. \quad (2)$$

Observing a censored observation  $(y_i, 0)$  is simply seeing an event that occurs with probability  $S_i(y_i | \theta)$ . In terms of the hazard function, the likelihood can be written

$$L(\theta | D) \propto \prod_{i=1}^n [h_i(y_i | \theta)]^{\delta_i} [S_i(y_i | \theta)]. \quad (3)$$

The contribution to the likelihood function for the  $i$ th individual is based on the observed data  $(y_i, \delta_i)$ . Denote it  $L(\theta | y_i, \delta_i)$ . We now show that

$$L(\theta | y_i, \delta_i) \propto [f_i(y_i | \theta)]^{\delta_i} [S_i(y_i | \theta)]^{1-\delta_i}.$$

To simplify the argument, we derive the contribution under the assumption that both the event time and censoring distributions are discrete. Let  $f_i(\cdot | \theta)$  and  $S_i(\cdot | \theta)$  be the discrete density and survivor function for  $T_i$ , and let  $g_i(\cdot)$  and  $G_i(\cdot)$  be the discrete density and survivor function for  $C_i$ . The likelihood contribution for individual  $i$  is

$$L(\theta | y_i, \delta_i) = \Pr(T_i = y_i, \delta_i = 1 | \theta)^{\delta_i} \Pr(C_i = y_i, \delta_i = 0)^{1-\delta_i}.$$

Noting that  $\delta_i = 1$  if and only if  $T_i \leq C_i$ ,

$$\begin{aligned} L(\theta | y_i, \delta_i) &= \begin{cases} \Pr(T_i = y_i, \delta_i = 1 | \theta), & \delta_i = 1 \\ \Pr(C_i = y_i, \delta_i = 0 | \theta), & \delta_i = 0 \end{cases} \\ &= \begin{cases} \Pr(T_i = y_i, T_i \leq C_i | \theta), & \delta_i = 1 \\ \Pr(C_i = y_i, T_i > C_i | \theta), & \delta_i = 0 \end{cases} \\ &= \begin{cases} \Pr(T_i \leq C_i | T_i = y_i, \theta) \Pr(T_i = y_i | \theta), & \delta_i = 1 \\ \Pr(T_i > C_i | C_i = y_i, \theta) \Pr(C_i = y_i | \theta), & \delta_i = 0 \end{cases} \\ &= \begin{cases} \Pr(y_i \leq C_i) \Pr(T_i = y_i | \theta), & \delta_i = 1 \\ \Pr(T_i > y_i | \theta) \Pr(C_i = y_i), & \delta_i = 0 \end{cases} \\ &= \begin{cases} [G_i(y_i) + g_i(y_i)] f_i(y_i | \theta), & \delta_i = 1 \\ S_i(y_i | \theta) g_i(y_i), & \delta_i = 0 \end{cases} \\ &\propto f_i(y_i | \theta)^{\delta_i} S_i(y_i | \theta)^{1-\delta_i}. \end{aligned}$$

The fourth equality uses the independence of  $C_i$  and  $T_i$  and the fact that censoring is uninformative, i.e., the censoring distribution does not depend on parameters of the event time. This argument can be modified to handle continuous cases.

In the remainder of this chapter we get our feet wet with one- and two-sample inference and a discussion of plotting the results of the analyses.

## 12.2 One-Sample Models

We begin with details of the three distributions that we most commonly use. We then discuss the likelihood function for one-sample data and the analysis for each of the three distributions.

### 12.2.1 Distributional Models

The simplest survival model is the Exponential distribution. The Exponential arises naturally as the waiting time between events in a Poisson process. A Poisson process counts the number of events that occur by time  $t$  under reasonable conditions discussed by Ross (2006). The name arises because for a fixed time, the number of events is Poisson, cf. Section 5.3.

EXAMPLE 12.2.1. *Exponential.* Let  $T|\theta \sim \text{Exp}(\theta)$ , then

$$f(t|\theta) = \theta e^{-\theta t}, \quad S(t|\theta) = e^{-\theta t} \quad t > 0.$$

The median,  $\tilde{t}$ , has  $0.5 = \exp(-\theta \tilde{t})$ . Solving, we get  $\tilde{t} = \log(2)/\theta = 0.69/\theta$ , which obviously depends on  $\theta$ . Other percentiles are found similarly. The hazard function is obtained as  $f(t|\theta)/S(t|\theta) = \theta e^{-\theta t}/e^{-\theta t} = \theta$ , which is constant in  $t$ . Thus, if the time to event has an Exponential distribution, the hazard of an event occurring is the same no matter what the time. This is called the *memoryless property* of the Exponential since the hazard of an event after, say, 100 years is the same as the hazard of the event at time 0.

The Weibull distribution is a generalization of the Exponential that incorporates a power transformation. If  $T \sim \text{Weib}(\alpha, \lambda)$ , then  $T^\alpha \sim \text{Exp}(\lambda)$ ; thus  $\alpha$  is the power that it takes to transform the Weibull into an Exponential. A Weibull with  $\alpha = 1$  is an Exponential.

EXAMPLE 12.2.2. *Weibull.* Define  $T|\alpha, \lambda \sim \text{Weib}(\alpha, \lambda)$  if

$$f(t|\alpha, \lambda) = \lambda \alpha t^{\alpha-1} e^{-\lambda t^\alpha}, \quad S(t|\alpha, \lambda) = e^{-\lambda t^\alpha}, \quad t > 0, \alpha > 0, \lambda > 0.$$

All percentiles of Weibulls are easy to find. The  $1 - \beta$  percentile of the Weibull is the value  $\gamma$  that satisfies  $\beta = \Pr(T > \gamma) = \exp[-\lambda \gamma^\alpha]$ . Solving, we obtain  $\gamma = [-\log(\beta)/\lambda]^{1/\alpha}$ . Of course  $\gamma$  depends on all of  $\beta$ ,  $\alpha$ , and  $\lambda$ . In particular, the median  $\tilde{t}$  is the time at which the survival probability is 0.50. It has  $0.50 = e^{-\lambda \tilde{t}^\alpha}$  and  $\tilde{t} = [\log(2)/\lambda]^{1/\alpha}$ . The mean is more complicated but we focus on medians in survival analysis. The hazard function is

$$h(t|\alpha, \lambda) = \frac{f(t|\alpha, \lambda)}{S(t|\alpha, \lambda)} = \lambda \alpha t^{\alpha-1}.$$

As a function of time, the hazard is increasing if  $\alpha > 1$ ; decreasing if  $\alpha < 1$ .

EXERCISE 12.1. Use Proposition B.4 to show that if  $T \sim \text{Weib}(\alpha, \lambda)$ , then  $T^\alpha \sim \text{Exp}(\lambda)$ .

Finally, we review our old friend the log-normal. If  $\log(T) \sim N(\mu, 1/\tau)$ , write  $T \sim LN(\mu, 1/\tau)$ .

EXAMPLE 12.2.3. *Log-Normal.* If  $T \sim LN(\mu, 1/\tau)$ , for  $t > 0$ ,

$$\begin{aligned} f(t|\mu, \tau) &= \frac{\sqrt{\tau}}{\sqrt{2\pi t}} \exp\left\{-\frac{\tau}{2}(\log(t) - \mu)^2\right\}, \\ S(t|\mu, \tau) &= 1 - \Phi[\sqrt{\tau}(\log(t) - \mu)], \end{aligned}$$

where  $\Phi(\cdot)$  is the cdf for the  $N(0, 1)$  distribution. The median  $\tilde{t}$  has  $0.50 = \Pr(T \leq \tilde{t}) = \Pr[\log(T) \leq \log(\tilde{t})]$ , so  $\log(\tilde{t}) = \mu$ , the center of symmetry of the  $N(\mu, 1/\tau)$  distribution. Obviously,  $\tilde{t} = e^\mu$ . The distribution of  $T$  is skewed, so the mean is not as useful as the median, and slightly more difficult to obtain. We won't use it. The hazard function does not have a simple form so we also don't write it. The  $\alpha$  percentile of the log-normal distribution is  $\gamma = e^{\mu + z_\alpha \sigma}$  where  $\sigma = 1/\sqrt{\tau}$  and  $z_\alpha$  is the  $\alpha$  percentile of the  $N(0, 1)$ . This occurs because  $\Pr[T \leq e^{\mu + z_\alpha \sigma}] = \Pr[\log(T) \leq \mu + z_\alpha \sigma] = \Pr[(\log(T) - \mu)/\sigma \leq z_\alpha] = \alpha$ .

**EXERCISE 12.2.** (a) Let  $W \equiv \log(T) \sim N(\mu, 1/\tau)$ . Derive the density, and survivor function for  $T$ . (b) In Exercise 4.8 we showed that the likelihood, and therefore the Bayesian analysis, does not depend on whether we transform the data before analyzing them. Show that this result remains true when the data are censored.

**EXERCISE 12.3.** Plot hazard, density, and survivor functions for the Weibull and log-normal distributions that have the following (median, 90th percentile) pairs: (1, 20), (20, 30), and (50, 70). Compare these Weibull and log-normal models by visual inspection. See Section 4 for information on plotting.

### 12.2.2 Posterior Analysis

We assume event times that are iid from  $f(\cdot | \theta)$  and subject to censoring. With data  $D = (y, \delta)$  where  $y = (y_1, \dots, y_n)'$  and  $\delta = (\delta_1, \dots, \delta_n)'$ , the likelihood function can be found from equation (12.1.2). With  $f_i(\cdot | \theta) = f(\cdot | \theta)$  etc.,

$$L(\theta | D) \propto \prod_{i=1}^n [f(y_i | \theta)]^{\delta_i} [S(y_i | \theta)]^{1-\delta_i}$$

or, from (12.1.3),

$$L(\theta | D) \propto \prod_{i=1}^n [h(y_i | \theta)]^{\delta_i} [S(y_i | \theta)].$$

With prior  $p(\theta)$ , the posterior is obtained as usual. Occasionally, the posterior is recognizable, as we shall see with the Exponential distribution. More often it is not recognizable, so we rely on simulations. Even when an analytical solution exists, it is often easier to use simulations. As usual, we obtain a posterior sample  $\{\theta^k : k = 1, \dots, m\}$  through Gibbs sampling. For any function of  $\theta$ , say  $\gamma = g(\theta)$ , we approximate the posterior  $p(\gamma | D)$  numerically by using the MC sample  $\{\gamma^k \equiv g(\theta^k) : k = 1, \dots, m\}$ .

*Censoring is a key feature in defining the likelihood and for obtaining a posterior sample. However, censoring plays no role in determining a prior or in interpreting the results once we have our posterior sample.*

### 12.2.3 Log-Normal Data

We begin with the log-normal distribution because we can dispose of it quickly. Analysis for the log-normal sampling model is similar to treatment of ordinary normal data. If we transform the data with logarithms, and there are no censored cases, the likelihood function takes the same form that it did in Section 5.2. Adding censored observations impacts the likelihood and the posterior but not the prior. Neither reference priors nor prior elicitation changes because of censoring. The SIR prior is still  $p(\mu, \tau) \propto 1/\tau$  but the posterior is more complicated with censoring. Our standard prior for this model is a normal distribution for  $\mu$  and an independent Gamma distribution for either  $\tau$  or  $\sigma$ . We sometimes specify a uniform distribution for  $\sigma$ . As illustrated in Sections 5.2 and 9.5, priors are induced from information on the median and a percentile of the distribution for  $T$ . Prior elicitations and posterior inferences about medians and other percentiles transform easily from the original scale to the log scale and back again. We defer further discussion of the log-normal distribution to Subsection 12.3.3. See Subsection 12.2.5 for computational methods.

### 12.2.4 Exponential Data

Consider data  $D = (y, \delta)$  where  $y_i = \min\{T_i, C_i\}$  and  $T_i \stackrel{iid}{\sim} \text{Exp}(\theta)$ . If the prior is taken to be  $\theta \sim \text{Gamma}(a, b)$ , then the posterior is

$$\begin{aligned} p(\theta | D) &\propto L(\theta | D)p(\theta) \\ &\propto \prod_{i=1}^n [\theta e^{-\theta y_i}]^{\delta_i} [e^{-\theta y_i}]^{1-\delta_i} \theta^{a-1} e^{-\theta b} \\ &= \theta^{a+n_u-1} e^{-\theta(b+\sum_{i=1}^n y_i)}, \end{aligned}$$

where  $n_u = \sum_{i=1}^n \delta_i$  is the number of uncensored observations. It follows that

$$\theta | D \sim \text{Gamma}(a + n_u, b + n\bar{y}). \quad (1)$$

While we focus on numerical approximations to posteriors, there is value in analytic results when they are available.

We now derive exact PIs for  $\theta$  when  $2(a + n_u)$  is an integer. From Exercise 5.19, when  $w \sim \text{Gamma}(a, b)$ , we have  $cw \sim \text{Gamma}(a, b/c)$ . Recall that  $\text{Gamma}(k/2, 1/2) \sim \chi_k^2$ . Since the posterior is equivalent to  $2(b + n\bar{y})\theta | D \sim \chi_{2(a+n_u)}^2$ , we can write

$$1 - \alpha = \Pr(\ell \leq c(y)\theta \leq u | D) = \Pr\left(\frac{\ell}{c(y)} \leq \theta \leq \frac{u}{c(y)} | D\right),$$

where  $c(y) = 2(b + n\bar{y})$  with  $\ell = \chi_{2(a+n_u)}^2(\alpha/2)$  and  $u = \chi_{2(a+n_u)}^2(1 - \alpha/2)$ , the lower and upper  $\alpha/2$  quantiles of the distribution. Thus,  $\ell/c(y) \leq \theta \leq u/c(y)$  is an exact  $100(1 - \alpha)\%$  PI for  $\theta$ . Of course, even when  $2(a + n_u)$  is not an integer an exact PI for  $\theta$  can always be determined by finding the appropriate percentiles of the posterior Gamma distribution computationally, cf. Section C.2.

How do we select the prior for  $\theta$ ? Our procedure is similar to Subsection 5.3.2 where a  $\text{Gamma}(a, b)$  prior was placed on a Poisson rate. Start by thinking about the median time to event,  $\tilde{t} = \log(2)/\theta$ . Suppose our best guess is  $\tilde{t}_0 = 10$  years. Then our best guess for  $\theta$  is  $\theta_0 \equiv \log(2)/\tilde{t}_0 = 0.069$ . Our  $\text{Gamma}(a, b)$  prior has mode  $\theta_0 = (a - 1)/b$ . Solving for  $a$  gives  $a = 1 + \theta_0 b = 1 + 0.069 b$ . What remains is to determine  $b$ .

To get  $b$  think about how large (or small) the median time to event  $\tilde{t}$  could be. If our expert is 95% sure that  $\tilde{t}$  is less than  $u = 20$  years, we can convert that into a probability statement about  $\theta$ . With  $\Pr[\tilde{t} \leq u] = 0.95$  and  $\tilde{t} = 0.69/\theta$ , we have  $0.95 = \Pr[0.69/\theta \leq u] = \Pr[\theta \geq 0.69/u] = \Pr[\theta \geq 0.0345]$  for  $u = 20$ . We are left to find  $b$  such that the  $\text{Gamma}(1 + 0.069 b, b)$  distribution has 0.0345 as its 5th percentile. We found  $b = 61$ , so  $a = 5.209$ .

### 12.2.5 WinBUGS for Censored Data

This subsection provides WinBUGS code for fitting the one-sample models. WinBUGS requires that the data be entered differently than  $D = (y, \delta)$ . With  $NA$  used to indicate missing data, transform  $(y, \delta)$  into  $(t, c)$  where for  $i = 1, \dots, n$

$$t_i = \begin{cases} y_i & \text{if } \delta_i = 1 \\ NA & \text{if } \delta_i = 0 \end{cases} \quad c_i = \begin{cases} 0 & \text{if } \delta_i = 1 \\ y_i & \text{if } \delta_i = 0 \end{cases}.$$

These define separate data vectors for uncensored and censored observations. In  $t$  you must use  $NA$  for each  $y_i$  that corresponds to a censored observation. The vector  $c$  must have the censoring time for all censored observations. Zeros work best for  $c_i$ s that correspond to noncensored cases.

To indicate that censored data are forthcoming, the distribution is identified as  $t \sim \text{ddist(theta)} \text{I}(a, b)$  where the numbers  $a$  and  $b$  provide information on censoring. *If data appear for  $t$  that are between  $a$  and  $b$ , the term  $\text{I}(a, b)$  is ignored.* If the data are not between  $a$  and  $b$ , an error message should ensue. However, if no data are available for  $t$ , it is assumed that the observation is from the specified distribution but censored as  $a < t < b$ . This general form allows the analysis of interval censored data. For the most common form of censoring, right censoring, use  $\text{I}(a, )$ , which indicates that  $a < t$ . For left censoring,  $\text{I}(, b)$  indicates that  $t < b$ .

For the Exponential model, calculating the median time to event and the survival probability at  $t = 5$ ,

```
model{
  for(i in 1:n){ t[i] ~ dexp(theta)I(c[i],) }
  theta ~ dgamma(a,b)
```

```

med <- 0.69/theta
S <- exp(-theta*tu)
}
list(n=100, tu=5, a=0.001, b=0.001)
t[ ] c[ ]
1      0
NA    2.5
more rows of data would be listed here
END

```

When WinBUGS sees the NA, it knows that the actual value for that individual is censored and generates an appropriate term for the likelihood. When it sees an actual data point for  $t$ , it ignores any information in the  $I(c[i],)$  statement and generates the appropriate term for the likelihood. (If  $c_i$  is larger than  $t_i = y_i$ , you should get an error message, so for uncensored observations use  $c_i = 0$  to be safe.)

### 12.2.6 Weibull Data

Let  $T_1, \dots, T_n$  be iid  $\text{Weib}(\alpha, \lambda)$ . If  $D = (y, \delta)$  is observed, the likelihood is

$$\begin{aligned} L(\alpha, \lambda | D) &= \prod_{i=1}^n [\lambda \alpha y_i^{\alpha-1}]^{\delta_i} \exp(-\lambda y_i^\alpha) \\ &= (\lambda \alpha)^{n_u} \left\{ \prod_{i=1}^n y_i^{\delta_i} \right\}^{\alpha-1} \exp\left(-\lambda \sum_{i=1}^n y_i^\alpha\right). \end{aligned}$$

Again  $n_u = \sum_{i=1}^n \delta_i$  is the number of uncensored observations and define  $v = -\sum_{i=1}^n \delta_i \log(y_i)$  and  $w(\alpha) = \sum_{i=1}^n y_i^\alpha$  so that

$$L(\alpha, \lambda | D) \propto \lambda^{n_u} e^{-\lambda w(\alpha)} \alpha^{n_u} e^{-\alpha v}. \quad (2)$$

This is not the functional form of a recognizable distribution unless  $\alpha$  is known, which it rarely is, so regardless of the prior we will not be able to find the posterior analytically. The problem is that  $w(\alpha)$  depends on  $\alpha$  in a complicated way. No prior can be placed on  $(\alpha, \lambda)$  that leads to a recognizable, analytically tractable joint posterior distribution. We must rely on simulations. The remainder of this subsection is devoted to identifying a prior.

The problem with specifying a prior on the model parameters is that there is no nice interpretation for  $\lambda$  when  $\alpha \neq 1$ . Moreover,  $\alpha$  is difficult to think about. Our idea is to pick two alternative parameters that are easy to think about and specify a joint prior for them. We then induce a prior on  $(\alpha, \lambda)$ . A natural choice for one alternative parameter is the median  $\tilde{t} = (\log(2)/\lambda)^{1/\alpha}$ . Two candidates for the second parameter are the  $1 - \beta$  percentile  $\gamma = [-\log(\beta)/\lambda]^{1/\alpha}$ , say  $1 - \beta = 0.90$ , or perhaps the 5-year survival rate,  $\eta \equiv e^{-\lambda 5^\alpha}$ . This technique presupposes that we can solve for the model parameters  $\alpha, \lambda$  in terms of  $\tilde{t}$  and either  $\gamma$  or  $\eta$ . Details of constructing such an informative prior can be culled from the discussion of constructing priors in Subsection 9.8. For now, we assume that we can elicit information on  $\alpha$  and  $\tilde{t}$ . Specifically, we assume best guesses  $\alpha_0$  and  $\tilde{t}_0$  as well as 95% upper bounds  $u_\alpha$  and  $u_{\tilde{t}}$ .

With  $\alpha > 0$ , take

$$\xi \equiv \log(\alpha) \sim N(c, d)$$

where  $c = \log(\alpha_0)$ . With  $\Pr[\alpha \leq u_\alpha] = 0.95$ , we get  $\Pr[\xi \leq \log(u_\alpha)] = 0.95$ , so standardizing gives

$$0.95 = \Pr\left\{[\xi - c]/\sqrt{d} \leq [\log(u_\alpha) - c]/\sqrt{d}\right\}.$$

Since 1.645 is the 95th percentile of the standard normal distribution, we have  $1.645 = [\log(u_\alpha) - c]/\sqrt{d}$ . Solving gives

$$d = [\log(u_\alpha) - c]^2 / 1.645^2.$$

If we have  $\alpha_0 = 1$  and  $u_\alpha = 5$ , then  $d = 0.957$ .

Of course, one might not be comfortable specifying an informative prior on  $\alpha$ . Recall that when  $\alpha = 1$ , the distribution is Exponential with a constant hazard, when  $\alpha < 1$  the hazard is decreasing, and for  $\alpha > 1$  it is increasing. A convenient reference prior takes  $\alpha_0 = 1$  and  $\log(\alpha) \sim N(0, d)$  with  $d$  large. Alternatively, one might pick  $\alpha \sim U(0, \bar{u})$  for some arbitrary large value of  $\bar{u}$ .

We specify an independent Gamma( $a, b$ ) prior for  $\lambda$ . The median is

$$\tilde{t} = (\log(2)/\lambda)^{1/\alpha},$$

so take the best guess  $\lambda_0$  to satisfy

$$\tilde{t}_0 = (\log(2)/\lambda_0)^{1/\alpha_0}$$

or

$$\lambda_0 = 0.69/\tilde{t}_0^{\alpha_0}.$$

With  $\lambda_0$  as the mode of the prior, we have  $a = 1 + \lambda_0 b$ .

To identify  $b$ , consider

$$\Pr[\tilde{t} \leq u_{\tilde{t}}] = 0.95.$$

Substitute the best guess for  $\alpha$  and write

$$0.95 = \Pr[(\log(2)/\lambda)^{1/\alpha_0} \leq u_{\tilde{t}}] = \Pr[0.69/u_{\tilde{t}}^{\alpha_0} \leq \lambda].$$

The 5th percentile of the prior on  $\lambda$  is  $0.69/u_{\tilde{t}}^{\alpha_0}$  and we proceed to find the  $\text{Gamma}(1 + \lambda_0 b, b)$  distribution that has this percentile. Larger values of  $u_{\tilde{t}}$  result in more diffuse priors on  $\lambda$ .

A flaw in this prior is assuming independence of knowledge about  $\alpha$  and  $\lambda$ . Clearly, we chose our prior on  $\lambda$  by taking  $\alpha = \alpha_0$ . We hope that if the prior on  $\lambda$  is sufficiently diffuse, this won't matter.

These are by no means the definitive priors for this problem. One could easily replace the log-normal prior on  $\alpha$  with a Gamma prior. None of these priors are conjugate.

**EXERCISE 12.4.** (a) For the Weibull model, with a log-normal prior on  $\alpha$  and a Gamma prior on  $\lambda$ , obtain the prior for  $(\alpha, \lambda)$  that has  $\tilde{t}_0 = 20$ ,  $\alpha_0 = 1$ ,  $\Pr[\alpha < 10] = 0.9$ , and  $\Pr[0.69/\lambda < 50 | \alpha = 1] = 0.95$ . (b) Use the following code to induce prior distributions on the median and the 20-year survival rates.

```
model{
  lambda ~ dgamma(a,b)
  alpha ~ dlnorm(c,d)
  # alpha ~ dgamma(c,d)
  # alpha ~ dunif(0,u)
  med <- pow(0.69/lambda,1/alpha)
  surv <- exp(-lambda*pow(20,alpha))
}
```

(c) Pick one of the alternative choices of prior mentioned above and modify the code to induce prior distributions on the median and 20-year survival rates. Compare these with the induced distributions from (b) and see how the induced prior on the median conforms with the specified prior information on the median under the Exponential model.

**EXERCISE 12.5.** Under the Weibull model, obtain the full conditional distribution for  $\lambda$  when the prior on  $\lambda$  is specified using a Gamma distribution independent of the prior on  $\alpha$ . Identify the distribution.

### 12.2.7 Prediction

Prediction is especially important in survival modeling. Let  $T_f$  denote a future survival time independent of the current data. The predictive density is

$$f(t|D) = \int f(t|\theta)p(\theta|D)d\theta.$$

The predictive survivor function is similarly obtained as

$$S(t|D) = \int S(t|\theta)p(\theta|D)d\theta.$$

From these one can find the predictive median or mean and probability intervals.

**EXAMPLE 12.2.4.** Consider  $T_f|\theta \sim \text{Exp}(\theta)$ . Using the posterior (1) and the fact that Gamma densities integrate to 1, the predictive survivor function is

$$\begin{aligned} S(t|D) &= \int_0^\infty e^{-\theta t} p(\theta|D)d\theta \\ &= \int_0^\infty e^{-\theta t} \frac{(b+n\bar{y}_.)^{a+n_u}}{\Gamma(a+n_u)} \theta^{a+n_u-1} e^{-\theta(b+n\bar{y}_.)} d\theta \\ &= \frac{(b+n\bar{y}_.)^{a+n_u}}{\Gamma(a+n_u)} \int_0^\infty \theta^{a+n_u-1} e^{-\theta(t+b+n\bar{y}_.)} d\theta \\ &= \frac{(b+n\bar{y}_.)^{a+n_u}}{\Gamma(a+n_u)} \frac{\Gamma(a+n_u)}{(t+b+n\bar{y}_.)^{(a+n_u)}} \\ &= \left( \frac{b+n\bar{y}_.}{t+b+n\bar{y}_.} \right)^{a+n_u}. \end{aligned}$$

The predictive density is the negative of the derivative of the predictive survivor function. Such clean analytical results are rare with censored data.

There are two ways to make predictive inferences in WinBUGS. The first approach is illustrated using the Exponential model in the following code:

```
model{
  for(i in 1:n){ t[i] ~ dexp(theta) I(c[i],) }
  theta ~ dgamma(a,b)
  tf ~ dexp(theta)
}
list(n=, a=, b=) # Sample size and prior parameters input by user
t[ ] c[ ]
# data lines
END
```

WinBUGS will recognize `tf` as a “future” observation and will sample from the full conditional distribution for it, which is precisely the Exponential distribution given the current iterate for  $\theta$ . Monitoring `tf` provides predictive inference.

The second way is to add at the end of the data lines a row with  $(t, c) = (\text{NA}, 0)$ , i.e., a missing response with  $c = 0$ . Now the data list is a matrix with  $n + 1$  rows where the first  $n$  rows contain the observed data. WinBUGS recognizes that the  $(n + 1)$ st observation is missing, and samples from the full conditional distribution for it, just as before. Predictive inference is obtained by monitoring `t[n+1]`. The same technique applies to other distributions.

### 12.2.8 Interval Censoring

We have mentioned interval censoring, let's look at it more carefully. Suppose we have event times

$$T_1, \dots, T_n | \theta \stackrel{iid}{\sim} f(t|\theta)$$

but each  $T_i$  is possibly censored in such a way that we would only observe that it falls between two numbers  $L_i < U_i$ . As before, let  $\delta_i$  denote the indicator random variable of noncensoring. Thus, the observed data are

$$\delta_i = \begin{cases} 1 & \text{if uncensored} \\ 0 & \text{if censored} \end{cases}$$

and  $T_i = t_i$  if  $\delta_i = 1$  or  $T_i \in (L_i, U_i)$  if  $\delta_i = 0$ . The standard “right” censoring scheme discussed earlier has  $L_i \equiv C_i$  and  $U_i \equiv \infty$ .

The likelihood function for data  $D$  is

$$L(\theta | D) = \prod_{i=1}^n [f(t_i | \theta)]^{\delta_i} \left[ \int_{L_i}^{U_i} f(t | \theta) dt \right]^{1-\delta_i}.$$

Note that if  $\delta_i = 0$ , all that is known is  $L_i < T_i < U_i$ .

**EXAMPLE 12.2.5.** Suppose the data are log-normal with parameters  $\mu$  and  $\tau$ , in other words,

$$\log(T_i) | \mu, \tau \stackrel{iid}{\sim} N(\mu, 1/\tau) \quad \text{or} \quad T_i | \mu, \tau \stackrel{iid}{\sim} LN(\mu, 1/\tau)$$

and use the prior

$$\mu \sim N(5, 100) \quad \perp \!\!\! \perp \quad \tau \sim \text{Gamma}(300, 3).$$

To specify this model in WinBUGS, we need to read in three data variables  $t[i]$ ,  $L[i]$ ,  $U[i]$  for each subject. The WinBUGS program is

```
model{
  for(i in 1:n){ t[i] ~ dlnorm(mu,tau) I(L[i], U[i]) }
  tau ~ dgamma(300, 3)
  mu ~ dnorm(5, 0.01)
  n <- # Sample size
}
t[] L[] U[]
# rows of data records
END
```

For uncensored observations,  $t[i] = t_i$ , the observed time to the event. We set the corresponding  $L[i] = 0$  in the data and we set  $U[i]$  to be a very large number. For data that are interval censored, the value of  $t[i]$  is missing and recorded as NA. If an observation is right censored at  $c_i$ , we have  $L_i = c_i$ ,  $U_i = \infty$ . Since we cannot set a value to infinity, use a value that is extremely large, say 1,000 times the value of the largest event time in the data. If the observation is left censored at  $c_i$ , set  $U_i = c_i$  and set  $L_i$  equal to 0.

**EXERCISE 12.6.** With right censoring the data are unambiguously defined by the vectors  $y$  and  $\delta$ . We now define the data for interval censoring as two vectors  $\xi$  and  $\eta$  where

$$\xi_i = \begin{cases} t_i & \text{if uncensored} \\ L_i & \text{if censored} \end{cases} \quad \text{and} \quad \eta_i = \begin{cases} 0 & \text{if uncensored} \\ U_i & \text{if censored} \end{cases}.$$

Observing that  $\delta_i = 1 - I_{(0,\infty)}(\eta_i)$ , write the likelihood in terms of  $\xi$  and  $\eta$ .

EXERCISE 12.7. Turnbull and Weiss (1978) reported data on 191 California high school boys who were asked, “When did you first use marijuana?” The data only noted ages in years so we regard the ages as interval censored. Thus age 18 is the interval [18,18.99] in our analysis. Twelve boys indicated use before a given age, so those data are left censored. In addition, 89 boys were still pristine, so were right censored. We regarded left and right censored observations for a given year as being in the middle of the year, thus a (left or right) censored observation for year 18 is taken as 18.5 years. See Hamburg et al. (1975) for details of the study.

The data would normally appear as three columns of numbers with 191 rows. All the data are censored, so  $t[i]$  would always be NA. In applying  $t[i] \sim dlnorm(\mu, \tau)$   $I(L[i], U[i])$ , the variables  $L[i]$  and  $U[i]$  are defined appropriately for left, right, or interval censoring. Here the data are presented differently and tricks are used to define the appropriate WinBUGS likelihood for a log-normal model.

```

model{
  for(i in 1:21){
    n[i] ~ dbin(p[i],r[i])
    r[i] <- n[i]
    p[i] <- phi((log(u[i]) - mu)/sigma) - phi((log(l[i]) - mu)/sigma)
  }
  mu ~ dnorm(0,0.001)
  tau ~ dgamma(0.001,0.001)I(0.001,)
  sigma <- 1/sqrt(tau)
  # Inference
  med <- exp(mu)
  ninetypct <- med*exp(1.28*sigma)
  surv <- 1-phi((log(16) - mu)/sigma)
  junk <- age[1]
}
list(mu=0, tau=1)
age[ ] n[ ] l[ ] u[ ]
10 4 10 10.99
11 12 11 11.99
12 19 12 12.99
12 2 12.5 100
13 24 13 13.99
13 15 13.5 100
13 1 0 13.5
14 20 14 14.99
14 24 14.5 100
14 2 0 14.5
15 13 13 13.99
15 18 15.5 100
15 3 0 15.5
16 3 16 16.99
16 14 16.5 100
16 2 0 16.5
17 1 17 17.5
17 6 17.5 100
17 3 0 17.5
18 1 0 18.5
18 4 18.5 100
END

```

(a) Give a careful expression of the likelihood function based on this code and argue why it is the correct likelihood. Be sure to argue why the value 100 used repeatedly in the data above is appropriate. (b) Alternatively, rewrite the data with 191 rows, a standard log-normal model, and show that the analysis is the same as for this code. The first four rows of the current data would become the first  $4 + 12 + 19 + 2$  rows of the augmented data. (c) Revise the code to add any additional inferences you might think interesting, and analyze the data. Also find the DIC for this model. (When we ran this, the program got *trapped*. If that happens, just click update to get the chain running again. We had to do this twice.) (d) Revise the code to give a Weibull model analysis of the same data. Analyze the data and obtain the DIC. (e) Compare inferences based on the two models. Which model would you prefer, or does it matter?

### 12.3 Two-Sample Data

With independent right censored samples from two populations  $i = 1, 2$ , the observations are written as  $y = \{y_{ij} = \min(T_{ij}, C_{ij}) : i = 1, 2, j = 1, \dots, n_i\}$  and the noncensoring indicators are given as  $\delta = \{\delta_{ij} = I(T_{ij} \leq C_{ij}) : i = 1, 2, j = 1, \dots, n_i\}$ . Again,  $D \equiv (y, \delta)$ . With samples from two populations, the task is to compare the populations' survival prospects. This involves comparing their survival curves, say,  $S_1(t)$  and  $S_2(t)$  or comparing the densities for the groups,  $f_1(t)$  and  $f_2(t)$ . The top row of Figure 12.1 illustrates this idea. Alternatively, one could compare the hazard functions  $h_1(t)$  and  $h_2(t)$ , perhaps by examining the hazard ratio  $h_1(t)/h_2(t)$ , as illustrated in the bottom row of Figure 12.1. The relative median  $\tilde{t}_1/\tilde{t}_2$  is another commonly used effect measure.

#### 12.3.1 Two-Sample Exponential Model

Consider two independent samples from Exponential distributions with parameters  $\theta_1$  and  $\theta_2$ . Each distribution has a constant hazard. This model works well in many reliability applications for mature units (units that have been in service for a while) that are not wearing out. If a refrigerator does not break early, it takes a long time to wear out. Many refrigerators are replaced for reasons other than breakage. For refrigerators that have survived several weeks and are not extremely old, an Exponential model seems reasonable.

In modeling time of survival after diagnosis with a disease, Exponential distributions would rarely be used. Often the hazard of death increases with chronic diseases and decreases when cures may occur, e.g., the hazard of death from cancer often decreases with time from treatment. The leukemia data below are an exception. They have been established in the literature to fit Exponential models quite well.

**EXAMPLE 12.3.1. Leukemia Data.** Feigl and Zelen (1965) present data on the survival times in weeks of patients who were diagnosed with leukemia. The patients were classified according to two characteristics of white blood cells referred to as  $AG+$  and  $AG-$ . The  $n_1 = 17$  times from diagnosis to death for the  $AG+$  group are: 65, 156, 100, 134, 16, 108, 121, 4, 39, 143, 56, 26, 22, 1, 1, 5, 65, and the  $n_2 = 16$  observations for the  $AG-$  group are: 56, 65, 17, 7, 16, 22, 3, 4, 2, 3, 8, 4, 3, 30, 4, 43. The histograms in Figure 12.2 show empirically that the patients in the  $AG+$  group tend to have better survival prospects than the  $AG-$  patients. There is no censoring, so  $y_{ij} = t_{ij}$  and  $\delta_{ij} = 1$  for all  $i$  and  $j$ .

We use an independent two-sample Exponential model with independent gamma priors on the rates

<u>AG Positive Group</u>	<u>AG Negative Group</u>
$T_{11}, \dots, T_{1n_1}   \theta_1 \stackrel{iid}{\sim} \text{Exp}(\theta_1)$	$\perp\!\!\!\perp$
$f(t   \theta_1) = \theta_1 e^{-\theta_1 t}$	$f(t   \theta_2) = \theta_2 e^{-\theta_2 t}$
$F(t   \theta_1) = 1 - e^{-\theta_1 t}$	$F(t   \theta_2) = 1 - e^{-\theta_2 t}$

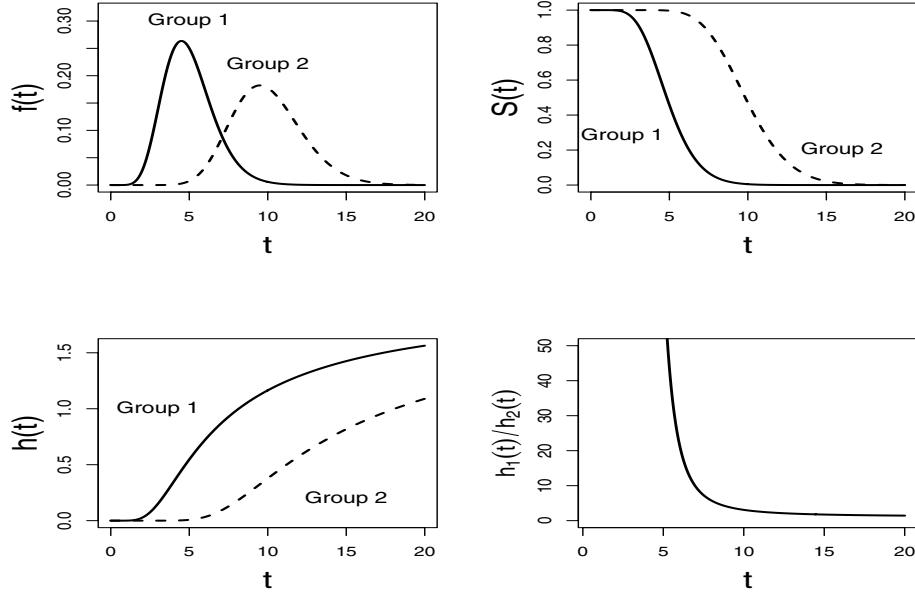


Figure 12.1: Illustration of effect measures for comparing two groups in survival analysis.

$$\begin{aligned} S(t|\theta_1) &= e^{-\theta_1 t} & S(t|\theta_2) &= e^{-\theta_2 t} \\ h(t|\theta_1) &= \theta_1 & h(t|\theta_2) &= \theta_2 \\ \theta_1 &\sim \text{Gamma}(a_1, b_1) & \perp\!\!\!\perp & \theta_2 \sim \text{Gamma}(a_2, b_2) \end{aligned}$$

For comparisons, the parameter of interest is  $\theta_2/\theta_1$ . It is the relative median time to death for group 1 versus group 2 since the median survival times are  $\log(2)/\theta_1$  and  $\log(2)/\theta_2$ . It is also the relative mean, since the mean survival times are  $1/\theta_1$  and  $1/\theta_2$ , and the relative hazard, since  $h(t|\theta_i) = \theta_i$ ,  $i = 1, 2$ . Here the ratio of hazards is constant over time. This is the simplest case of proportional hazards, a subject discussed in Section 13.2.

With no censored observations in the Feigl and Zelen data, denote the two observed samples  $t_1 = (t_{11}, \dots, t_{1n_1})$  and  $t_2 = (t_{21}, \dots, t_{2n_2})$ . The likelihood function is

$$\begin{aligned} L(\theta|t_1, t_2) &= \prod_{j=1}^{n_1} f(t_{1j}|\theta_1) \prod_{k=1}^{n_2} f(t_{2k}|\theta_2) \\ &= \prod_{j=1}^{n_1} \theta_1 e^{-\theta_1 t_{1j}} \prod_{k=1}^{n_2} \theta_2 e^{-\theta_2 t_{2k}} \\ &= \theta_1^{n_1} e^{-\theta_1(n_1 \bar{t}_1)} \theta_2^{n_2} e^{-\theta_2(n_2 \bar{t}_2)} \end{aligned}$$

where  $\bar{t}_i = \sum_{j=1}^{n_i} t_{ij}/n_i$  is the mean time to death for group  $i$ . With independent  $\text{Gamma}(a_i, b_i)$  priors, we obtain the joint posterior density

$$p(\theta_1, \theta_2 | t_1, t_2) \propto \theta_1^{a_1+n_1-1} e^{-\theta_1(b_1+n_1 \bar{t}_1)} \theta_2^{a_2+n_2-1} e^{-\theta_2(b_2+n_2 \bar{t}_2)},$$

which is the product of two independent gamma kernels, thus

$$\theta_1 | t_1 \sim \text{Gamma}(a_1 + n_1, b_1 + n_1 \bar{t}_1) \quad \perp\!\!\!\perp \quad \theta_2 | t_2 \sim \text{Gamma}(a_2 + n_2, b_2 + n_2 \bar{t}_2).$$

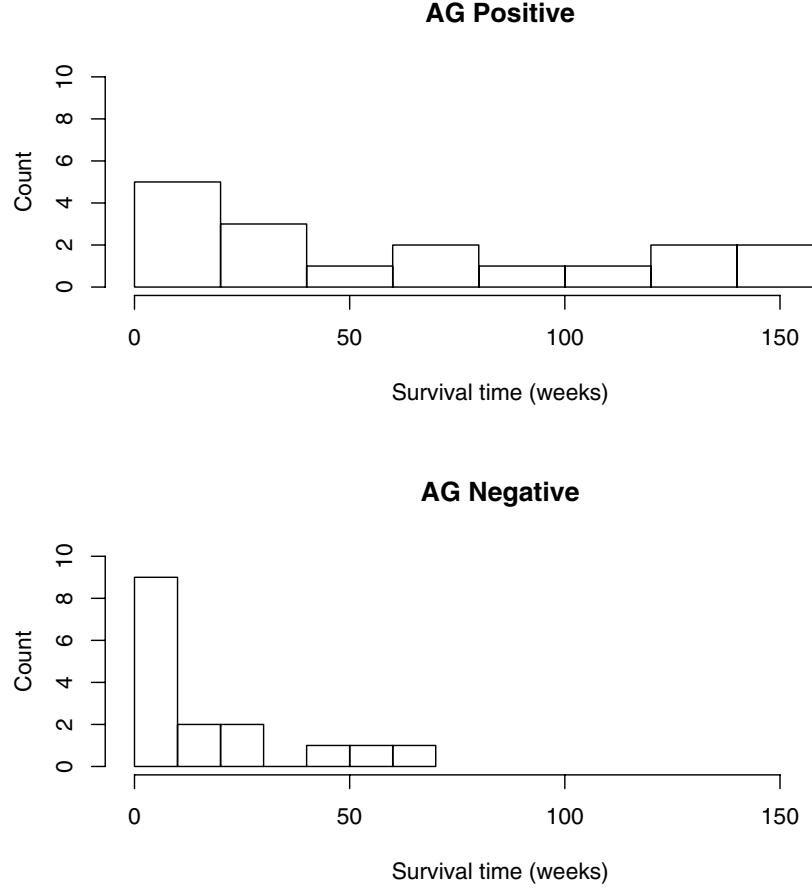


Figure 12.2: *Histograms of the leukemia data.*

As in the one-sample case, we focus mainly on numerical approximations to posteriors; however, we still develop analytic results for illustrative purposes. We derive an exact PI formula for  $\theta_2/\theta_1$  when  $2(a_i + n_i)$  is an integer for  $i = 1, 2$ . Recall that

$$2(b_1 + n_1 \bar{t}_1) \theta_1 | t_1 \sim \chi^2_{2(a_1 + n_1)} \quad \perp \quad 2(b_2 + n_2 \bar{t}_2) \theta_2 | t_2 \sim \chi^2_{2(a_2 + n_2)}.$$

The ratio of two independent  $\chi^2$  random variables, divided by their degrees of freedom, has an  $F$  distribution with corresponding numerator and denominator degrees of freedom, so given the data

$$\frac{(b_2 + n_2 \bar{t}_2) \theta_2}{a_2 + n_2} \Big/ \frac{(b_1 + n_1 \bar{t}_1) \theta_1}{a_1 + n_1} \equiv c(t_1, t_2) \frac{\theta_2}{\theta_1} \sim F_{2(a_2 + n_2), 2(a_1 + n_1)}.$$

We can now write

$$1 - \alpha = \Pr \left( \ell \leq c(t_1, t_2) \frac{\theta_2}{\theta_1} \leq u \mid t_1, t_2 \right) = \Pr \left( \frac{\ell}{c(t_1, t_2)} \leq \frac{\theta_2}{\theta_1} \leq \frac{u}{c(t_1, t_2)} \mid t_1, t_2 \right),$$

where  $\ell$  and  $u$  are the lower and upper  $\alpha/2$  quantiles of the above  $F$  distribution. Thus  $(\ell/c(t_1, t_2), u/c(t_1, t_2))$  is an exact  $100(1 - \alpha)\%$  PI for  $\theta_2/\theta_1$ . If  $2(a_i + n_i)$  is moderately large, say 10 or more, we round to the nearest integer. Alternatively, the method of composition can be used to simulate from the posterior distribution of  $\theta_2/\theta_1$  and does not require integer degrees of freedom.

This analysis is similar to the two-sample Poisson of Subsection 5.3.4. We start with independent Gamma priors and we end up with independent Gamma posteriors. The likelihood functions for Exponential and Poisson data have the same form, despite the fact that the sampling models are decidedly different: Poisson data are counts while Exponential responses are continuous. The Poisson distribution counts the number of events occurring within a certain amount of time and the exponential distribution measures the amount of time it takes for events (counts) to occur.

With censoring, the results change only slightly. Replace  $a_i + n_i$  with  $a_i + n_{iu}$  where  $n_{iu}$  is the number of uncensored observations in group  $i$ . Also,  $\bar{t}_i$  is now the average of all the censoring times together with all the uncensored observations.

**EXAMPLE 12.3.2. *Leukemia Data Continued.*** We construct a prior without the benefit of an expert so as to illustrate the method. Readers can judge our prior for themselves. Our best guess for  $\tilde{t}_2$ , the median survival time of AG– patients, is  $\tilde{t}_{2,0} = 20$  weeks and we believe that the median is greater than 5 weeks with 95% certainty. With  $\tilde{t}_{2,0} = \log(2)/\theta_{2,0} = 20$ , our prior point estimate of  $\theta_2$  is  $\theta_{2,0} = 0.69/20 = 0.0345$ . Also,

$$0.95 = \Pr[\tilde{t}_2 > 5] = \Pr[0.69/\theta_2 > 5],$$

so  $\Pr[\theta_2 < 0.138] = 0.95$ . We construct a Gamma prior for  $\theta_2$  that has a mode of  $(a_2 - 1)/b_2 = 0.0345$  and has 95th percentile equal to 0.138. The prior is  $\theta_2 \sim \text{Gamma}(2.31, 37.95)$ . For  $\theta_1$  we use a  $\text{Gamma}(1.53, 26.4)$ , which has mode 0.02 and 95th percentile 0.15. These correspond to a prior guess of  $\tilde{t}_{1,0} = 34.5$  weeks and a 5th percentile of 4.6 weeks. R code for finding  $\theta_2$ 's hyperparameters  $a_2$  and  $b_2$  follows.

```
b2 <- seq(0.001, 50, 0.011)
a2 <- 0.0345*b2 + 1
cbind(a2,b2,qgamma(0.95,a2,b2)) # Look for 0.138
```

A reference prior is the Jeffreys' prior  $p(\theta_1, \theta_2) \propto 1/(\theta_1 \theta_2)$ , which we could approximate with independent  $\text{Gamma}(0.001, 0.001)$  priors.

**EXERCISE 12.8.** Using our informative prior and the approximate Jeffreys prior, perform the following tasks for the leukemia data without using simulations (but possibly programming exact computations).

- (a) Obtain the joint posterior density for  $(\theta_1, \theta_2)$ .
- (b) Obtain an approximate 95% PI for the relative median survival time  $\theta_2/\theta_1$ , using the appropriate  $F$  distribution.
- (c) Obtain the posterior mean, median, mode, and (approximate) 95% PIs for  $\tilde{t}_1 = 0.69/\theta_1$  and  $\tilde{t}_2 = 0.69/\theta_2$ .
- (d) Compute the posterior probability that the relative median is greater than two.

**EXAMPLE 12.3.3. *Leukemia Data.*** We now illustrate the use of WinBUGS to perform the calculations of Exercise 12.8. The variables  $S[1]$  and  $S[2]$  in the following code are the 24-week (approximately 6-month) probabilities of survival for the two groups.

```
model
{
  for(i in 1:n[1]) {t1[i] ~ dexp(theta1)}
  for(i in 1:n[2]) {t2[i] ~ dexp(theta2)}
  theta1 ~ dgamma(a1,b1)
  theta2 ~ dgamma(a2,b2)
  median1 <- log(2)/theta1
  median2 <- log(2)/theta2
  relmedian <- theta2/theta1
```

Table 12.1: *WinBUGS output for leukemia data.*

node	mean	sd	2.5%	median	97.5%
median1	43.030	10.560	27.040	41.490	68.060
median2	13.030	3.234	8.124	12.550	20.650
relmedian	3.495	1.202	1.718	3.303	6.380
S[1]	0.668	0.062	0.541	0.670	0.783
S[2]	0.272	0.082	0.129	0.266	0.447
Sdiff	0.396	0.103	0.1834	0.400	0.585
theta1	0.017	0.004	0.010	0.017	0.026
theta2	0.056	0.013	0.034	0.055	0.085

Table 12.2: *Comparison of posterior medians and 95% PIs for 3 sets of priors in the leukemia analysis.*

Prior	$\theta_1$	$\theta_2$	med1	med2	$\theta_2/\theta_1$
Informative	0.017 (.010, .026)	0.055 (.034, .085)	41.5 (27.0, 68.0)	12.6 (8.1, 20.7)	3.3 (1.7, 6.4)
Diffuse Gamma	0.016 (.009, .025)	0.055 (.032, .087)	44.2 (28.3, 74.7)	12.7 (8.0, 21.7)	3.5 (1.7, 7.0)
Diffuse Uniform	0.016 (.009, .025)	0.055 (.032, .087)	44.2 (28.3, 74.7)	12.7 (8.0, 21.7)	3.5 (1.7, 7.0)

```

S[1] <- exp(-24*theta1)
S[2] <- exp(-24*theta2)
Sdiff <- S[1]-S[2]
}
list(n=c(17,16),
a1=, a2=, b1=, b2=, #From Example 12.3.2 or approximate Jeffreys
t1=c(65,156,100,134,16,108,121,4,39,143,56,26,22,1,1,5,65),
t2=c(56,65,17,7,16,22,3,4,2,3,8,4,3,30,4,43))
list(theta1=0.05,theta2=0.02)

```

Table 12.1 provides the WinBUGS output when using the informative priors from Example 12.3.2. The prior and posterior distributions for  $\theta_1$  and  $\theta_2$  are plotted in Figure 12.3. The median time to death for the AG+ group is estimated as 41.5 weeks with 95% PI (27, 68), much higher than for the AG- group with estimate 12.6 weeks and 95% PI (8, 21). The relative median  $\theta_2/\theta_1$  is roughly 3.3 and probably between 1.7 and 6.4. About 2/3 of AG+ patients will live at least 24 weeks, whereas only about 1/4 of AG- patients will live 24 weeks or longer. The PIs do not overlap. The difference in the 24-week survival probabilities is about 0.4 with a PI that is clearly positive.

A sensitivity analysis was conducted using the informative prior and two sets of diffuse priors:  $\theta_1, \theta_2 \stackrel{iid}{\sim} \text{Gamma}(0.001, 0.001)$  and  $\theta_1, \theta_2 \stackrel{iid}{\sim} U(0, 1000)$ . The Gamma priors have mean = 1 and variance 1,000. Both sets are diffuse since the leukemia survival times range from 2 to 156 weeks. Results appear in Table 12.2 along with some from the previous analysis. Even with the relatively small sample sizes, the different priors do not give appreciably different results.

**EXERCISE 12.9.** Consider survival data that are modeled as independent Exponential samples without censoring. (a) Using the Jeffreys prior for  $\theta_1$  and  $\theta_2$ , derive the analytical form of the joint predictive density for  $(y_{1f}, y_{2f})$ , a pair of future observations that are conditionally independent of the data and each other given  $(\theta_1, \theta_2)$ . Characterize it as best you can. (b) For the leukemia data, augment the WinBUGS code with informative priors for the  $\theta_i$ s to numerically simulate the joint predictive density and also the predictive density of  $y_{1f} - y_{2f}$ , and the predictive probabilities that

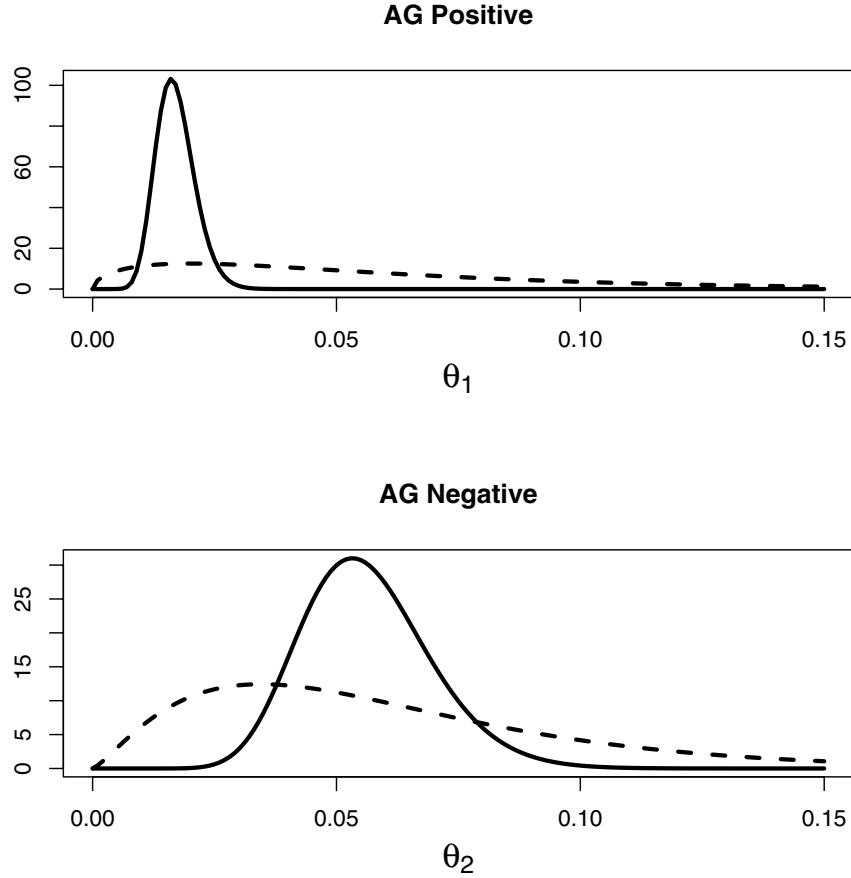


Figure 12.3: Prior (dashed lines) and posterior (solid lines) distributions for  $\theta_1$  and  $\theta_2$ .

$y_{1f} > y_{2f}$  and  $y_{1f} > y_{2f} + 10$ . (c) Obtain these same predictive inferences using the approximate Jeffreys prior and compare.

### 12.3.2 Two-Sample Weibull Model

Consider two independent samples

$$\begin{aligned} t_{11}, \dots, t_{1n_1} | \alpha_1, \lambda_1 &\stackrel{iid}{\sim} \text{Weib}(\alpha_1, \lambda_1) \\ t_{21}, \dots, t_{2n_2} | \alpha_2, \lambda_2 &\stackrel{iid}{\sim} \text{Weib}(\alpha_2, \lambda_2). \end{aligned}$$

These may be times it takes for brand A and brand B light bulbs to burn out. Define  $\alpha = (\alpha_1, \alpha_2)$  and  $\lambda = (\lambda_1, \lambda_2)$ , and assume that censored data  $D$  are observed. Similar to the one-sample case of Subsection 12.2.6, compute summary measures  $n_{iu} = \sum_{j=1}^{n_i} \delta_{ij}$ ,  $v_i = -\sum_{j=1}^{n_i} \delta_{ij} \log(y_{ij})$ , and  $w_i(\alpha_i) = \sum_{j=1}^{n_i} y_{ij}^{\alpha_i}$ . Using independence and the one-sample likelihood (12.2.2), we obtain the two-sample likelihood function

$$L(\alpha, \lambda | D) \propto \prod_{i=1}^2 \lambda_i^{n_{iu}} e^{-\lambda_i w_i(\alpha_i)} \alpha_i^{n_{iu}} e^{-\alpha_i v_i}. \quad (1)$$

This is just the product of the two one-sample likelihoods. For any of the priors discussed in Subsection 12.2.6, we use independent versions for each sample. The joint posterior is the product of independent posteriors for the two  $(\alpha_i, \lambda_i)$  pairs. We rely on Gibbs sampling for posterior inferences.

Below we provide WinBUGS code for specifying a two-sample Weibull model for handling the leukemia data (lists of the data and initial values are omitted).

```
model{
  for(i in 1:n[1]) {t1[i] ~ dweib(alpha[1], lambda[1])}
  for(i in 1:n[2]) {t2[i] ~ dweib(alpha[2], lambda[2])}
  lambda[1] ~ dgamma(a[1], b[1])
  alpha[1] ~ dlnorm(c[1], d[1])
  lambda[2] ~ dgamma(a[2], b[2])
  alpha[2] ~ dlnorm(c[2], d[2])
  med[1] <- pow(log(2)/lambda[1], 1/alpha[1])
  med[2] <- pow(log(2)/lambda[2], 1/alpha[2])
  relmedian <- med[1]/med[2]
  S[1] <- exp(-pow(24, alpha[1])*lambda[1])
  S[2] <- exp(-pow(24, alpha[2])*lambda[2])
}
```

**EXERCISE 12.10.** *Leukemia Data.* Analyze the leukemia data using the two-sample Weibull model. You must first select values for the prior parameters  $a_i, b_i, c_i$ , and  $d_i$ , for  $i = 1, 2$ . (a) For  $\lambda_1$  and  $\lambda_2$  use the Gamma priors that were constructed in Example 12.3.2 under Exponential sampling models for the data. Use  $LN(0, d)$  priors for the  $\alpha$ s, where you are 95% sure that  $\alpha_i \leq 10$ . (b) Now use  $\text{Gamma}(0.001, 0.001)$  priors for the  $\lambda$ s and use the prior in (a) for  $\alpha$ s. (c) Perform a sensitivity analysis by varying the choice of  $d$  and summarize your results in a table.

### 12.3.3 Two-Sample Log-Normal Model

Log-normal survival data were introduced in Example 12.2.3, which presented the density, survivor function, and median. As discussed in Subsection 12.2.3 and illustrated in Subsection 5.2, priors can be elicited on the original scale and transformed to the log scale and inferences on the log scale can be transformed easily to the original scale. The standard improper reference prior for a sample of normal or log-normal data is  $p(\mu, \tau) \propto 1/\tau$ .

The primary difference between two-sample log-normal survival analysis and the analysis of Subsection 5.2.5 is the introduction of censoring. With survival data from two independently sampled populations,  $LN(\mu_1, 1/\tau_1)$  and  $LN(\mu_2, 1/\tau_2)$ , just take natural logs of the observations and treat them as two-sample censored normal data. Censoring complicates the likelihood function and makes posterior distributions analytically intractable, regardless of the choice of prior. Fortunately, posterior simulation is not difficult. Elicitation of independent informative priors for the two groups follows exactly as in Subsection 5.2.5. Otherwise, the SIR prior is  $p(\mu_1, \mu_2, \tau_1, \tau_2) \propto 1/(\tau_1 \tau_2)$ , which is approximated by

$$\mu_1, \mu_2 \stackrel{iid}{\sim} N(0, 0.000001) \quad \perp \quad \tau_1, \tau_2 \stackrel{iid}{\sim} \text{Gamma}(0.001, 0.001).$$

**EXERCISE 12.11.** Analyze the leukemia data using the two-group log-normal model. (a) Use a proper approximation to the SIR prior. (b) Do your best to incorporate the prior information that was already used in this chapter for analyzing the leukemia data to place informative priors on  $\mu_1$  and  $\mu_2$ . Keep in mind that this prior was created only for illustration with no expert input.

Table 12.3: *Breast cancer data.*

t []	c []	h []	t []	c []	h []	t []	c []	h []
19	0	1	56	0	1	22	0	2
25	0	1	57	0	1	23	0	2
30	0	1	61	0	1	38	0	2
34	0	1	66	0	1	42	0	2
37	0	1	67	0	1	73	0	2
46	0	1	74	0	1	77	0	2
47	0	1	78	0	1	89	0	2
51	0	1	86	0	1	115	0	2
tt []	cc []	hh []	tt []	cc []	hh []	tt []	cc []	hh []
NA	122	1	NA	141	1	NA	156	1
NA	123	1	NA	143	1	NA	162	1
NA	130	1	NA	148	1	NA	164	1
NA	130	1	NA	151	1	NA	165	1
NA	133	1	NA	152	1	NA	182	1
NA	134	1	NA	153	1	NA	189	1
NA	136	1	NA	154	1	NA	144	2

EXERCISE 12.12. Table 12.3 contains censored survival times in months for 45 women with breast cancer presented in a WinBUGS format:  $t[]$  and  $c[]$ . The women all tested negative for axillary lymph node involvement by the standard method but were given an additional test for lymph node involvement: an immunohistochemical response  $h[]$  that takes the value 1 if the outcome was negative and 2 if it was positive. For computational reasons, the data have been divided into uncensored and censored observations. In the data, censoring only occurs after 10 years (120 months). We presume that it was impossible to get censored before 120 months. (Presumably, no one is lost to the study and censoring only occurs because the study is winding down. If people lost to the study within 10 years were excluded, it would violate our assumption of independent, noninformative censoring.) Of the 45 patients, 36 were negative and 9 were positive. Over half the subjects died and 8 of 9 people with positive immunohistochemical response died. The data originally come from Sedmark et al. (1989) and are also analyzed in Klein and Moeschberger (2003).

In constructing the WinBUGS likelihood we divided the 24 uncensored observations from the 21 censored ones but with the parameters of the distributions identical for the two parts. We did this because we got WinBUGS errors when we tried to combine all the data into a single loop. We believe the errors are due to WinBUGS having trouble generating initial values for a censored distribution. We used the actual censoring times as our initial values for the corresponding missing survival times.

```

model{
# Likelihood
for(i in 1:24){ t[i] ~ dlnorm(mu[h[i]],tau[h[i]])I(c[i],) }
for(i in 1:21){ tt[i] ~ dlnorm(mu[hh[i]], tau[hh[i]])I(cc[i],) }
# Prior
for(i in 1:2){
  mu[i] ~ dnorm(0,0.001)
  tau[i] ~ dgamma(0.001,0.001)
# Inference
med[i] <- exp(mu[i])
surv[i] <- 1- phi((log(120) - mu[i])/sigma[i])
sigma[i] <- 1/sqrt(tau[i])
}

```

```

}
relmed <- med[1]/med[2]
survdiff <- surv[1] - surv[2]
}
list(mu=c(0,0), tau=c(1,1), tt=c(122,123,130,130,133,134,136,
141,143,148,151,152,153,154,156,162,164,165,182,189,144))
t[ ] c[ ] h[ ]
19 0 1
... # Additional data lines
115 0 2
END
tt[ ] cc[ ] hh[ ]
NA 122 1
... # Additional data lines
NA 144 2
END

```

(a) Analyze the data with our two-sample log-normal survival model code. Modify the code to include additional inferences that you find interesting. Obtain the DIC value. (b) Modify the code to analyze the data under the assumption that  $\tau_1 = \tau_2$ . Compare some key inferences to see how different they are from part (a). Obtain the DIC value and decide which model is preferable. (c) Reanalyze the data using the Weibull model with (i) different values for  $(\alpha, \lambda)$  for the two populations and (ii) distinct values for  $\lambda$  but the same value for  $\alpha$ . Give inferences in both cases. Obtain DIC values. Which model is preferable? (d) Compare all models and decide on a preferred model.

## 12.4 Plotting Survival and Hazard Functions

As discussed in Subsection 12.1.1, primary objects of inference for time to event data are the survival curve and the hazard function. To be displayed effectively, information about these quantities should be plotted. Consider a general parametric model for survival data that depends on a vector of parameters  $\theta$  and a vector of covariate information  $x$ . For a one-sample problem,  $x$  never changes and can be ignored. For a two-sample problem,  $x$  identifies the two groups and  $\theta$  contains parameters for both groups. More complicated models involving  $x$  are discussed in Chapter 13.

To estimate  $S(t|x, \theta)$  for all  $t > 0$ , use MCMC methods to simulate  $\theta^1, \dots, \theta^m$  from the posterior distribution of  $\theta$ . For each  $\theta^k$  compute  $S(t|x, \theta^k)$ . Since we cannot evaluate the survival function at all  $t > 0$ , evaluate it over a fine grid, perhaps  $t = 0, 0.01, 0.02, \dots, T_*$ , where  $T_*$  is just bigger than the largest observed time in the data. (Some tuning may be needed to determine an appropriate  $T_*$  and reasonable spacing between grid points.) The left panel of Figure presents the first 20 out of 50,000 sampled curves for the  $AG+$  group from a two-sample Exponential analysis of the leukemia data.

The posterior mean of the survival and hazard functions are approximated by

$$\hat{S}(t|x, \theta) \doteq \frac{1}{m} \sum_{k=1}^m S(t|x, \theta^k) \quad \text{and} \quad \hat{h}(t|x, \theta) \doteq \frac{1}{m} \sum_{k=1}^m h(t|x, \theta^k).$$

The right panel of Figure presents posterior mean estimates of the survival functions for the leukemia study. Clearly the  $AG-$  group has substantially worse survival prospects than  $AG+$  patients. From  $\{S(t|x, \theta^1), \dots, S(t|x, \theta^m)\}$  we can calculate the posterior median and other percentiles of the survival function at each grid point. By interpolation between  $t$  values of the 2.5 and 97.5 percentiles, we can plot a 95% pointwise probability band for the survival function. For the leukemia study, these bands and the posterior medians are plotted for each group in Figure . We see that survival among  $AG+$  patients is substantially extended compared to  $AG-$  patients because the bands do not overlap. Construction of the plots in Figures and is discussed in Section C.5.

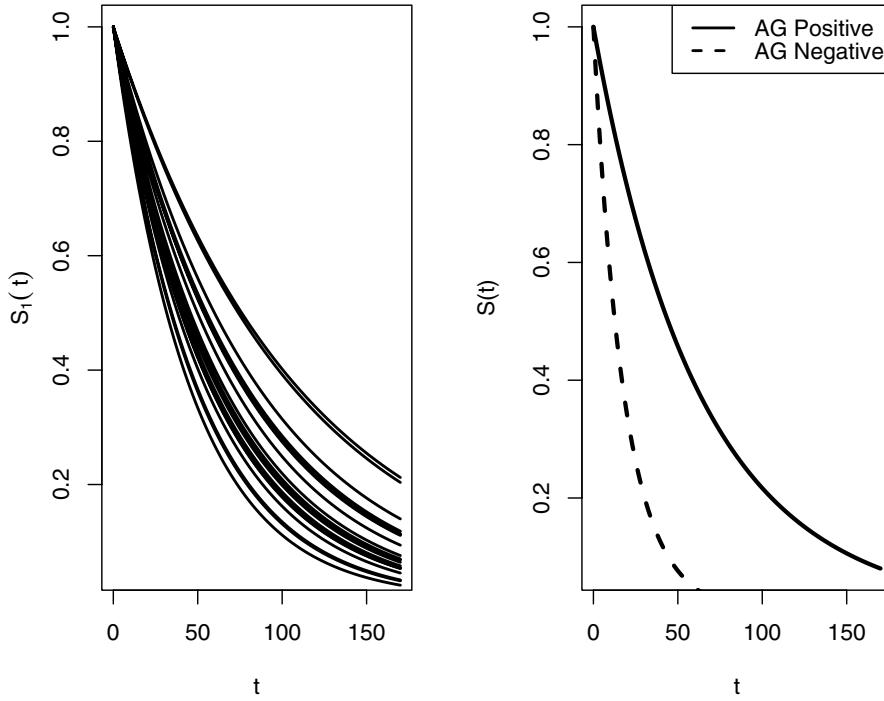


Figure 12.4: 20 simulated values of  $S_1(t)$  for the AG positive group (left panel), and the posterior means based on 50,000 simulated survival functions (right panel).

Estimating hazard functions for the leukemia data with Exponential distributions is easy because the hazard functions are constant in  $t$ . For the leukemia study, the hazard ratio  $\theta_2/\theta_1$  has posterior median 3.3 and 95% interval (1.7, 6.4). The estimated hazard of death at any time  $t$  for the AG- group is 3.3 times that for the AG+ group.

Generally, the hazard of event occurrence is  $h(t|x, \theta) = f(t|x, \theta)/S(t|x, \theta)$ , and changes with  $t$ . Estimation proceeds as outlined for survival curves. Using each  $\theta^k$ , evaluate  $h(t|x, \theta^k) = f(t|x, \theta^k)/S(t|x, \theta^k)$  on a grid of  $t$  values. For each  $t$ , compute the mean, median, and percentiles from  $\{h(t|x, \theta^1), \dots, h(t|x, \theta^m)\}$  and interpolate between  $t$  values.

Although it might be convenient to simply plug the posterior mean of  $\theta$  into  $S(t|x, \theta)$  or  $h(t|x, \theta)$ , typically the result is not the posterior mean of  $S(t|x, \theta)$  or  $h(t|x, \theta)$ . In other words, typically  $E[S(t|x, \theta)|D] \neq S(t|x, E[\theta|D])$  and  $E[h(t|x, \theta)|D] \neq h(t|x, E[\theta|D])$ . Moreover, typically  $\sum_{k=1}^m S(t|x, \theta^k)/m \neq S(t|x, \hat{\theta})$  for  $\hat{\theta} = \sum_{k=1}^m \theta^k/m$ . This “plug-in” approach is easier to compute and may in some cases provide a decent approximation but it does not yield easily interpretable Bayesian estimates of the survival or hazard functions.

On the other hand, a plug-in approach based on posterior percentiles works more often. If  $S(t|x, \theta)$  or  $h(t|x, \theta)$  is strictly monotone in a scalar parameter  $\theta$ , the posterior median of  $\theta$ , say  $\tilde{\theta} \equiv \tilde{\theta}(D)$ , should have  $S(t|x, \tilde{\theta})$  or  $h(t|x, \tilde{\theta})$  as their posterior medians. Similar results would hold for other percentiles. However, if  $\theta$  is a vector or if the functions are not strictly monotone, empirical percentiles of the  $S(t|x, \theta^k)$ s and the  $h(t|x, \theta^k)$ s need to be found for each grid value of  $t$ .

**EXERCISE 12.13. Leukemia data.** Give plots of survival and hazard functions for the two groups in the leukemia study using (a) the log-normal model, (b) the Weibull model, and (c) the

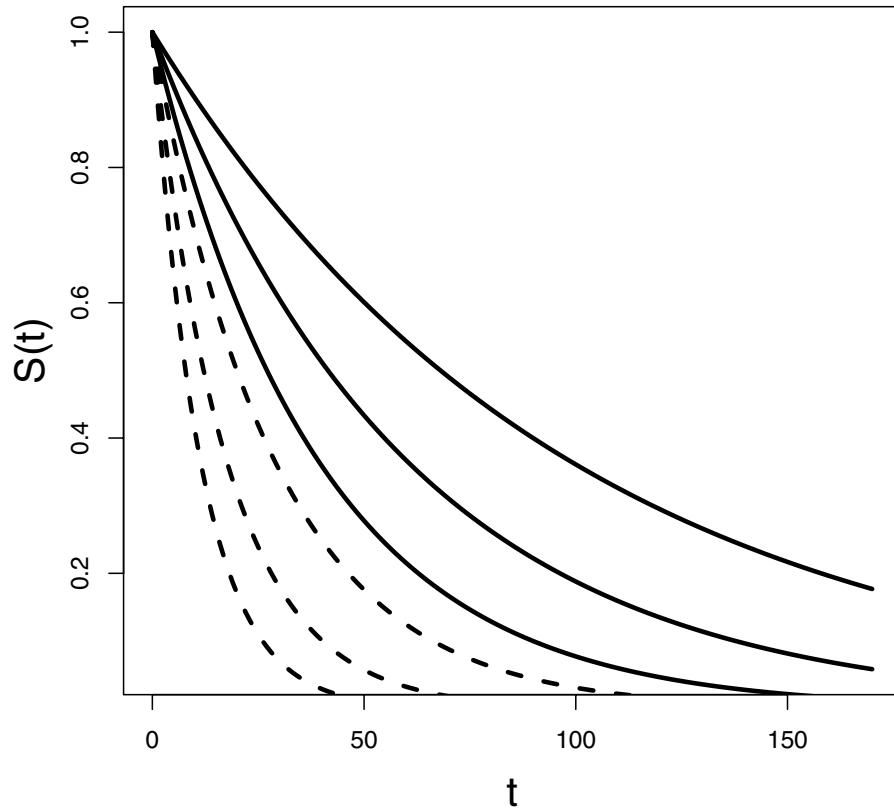


Figure 12.5: Posterior median and 95% pointwise probability bands for the AG positive (solid lines) and negative (dashed lines) groups.

Exponential model. Use relatively noninformative priors and obtain the DIC statistic in each case so that you can select a preferred model among these three candidates.

**EXERCISE 12.14. *Abortion in Dairy Cattle.*** Analyze the two-sample cow abortion data discussed in Section 1.4 using (a) log-normal and (b) Weibull models. Group membership is determined by infection status of an abortion-causing parasite called *Neospora caninum*. Give plots of survivor functions and hazard functions. Obtain the DIC statistic for each model and choose the preferable one.

---

## Chapter 13

---

# Time to Event Regression

---

We now introduce covariates into a time to event analysis via the *accelerated failure time (AFT)* model and the *proportional hazards (PH)* model. As in Chapter 12, we have a sample of data of the form  $D = (y, \delta)$  with  $y = (y_1, \dots, y_n)'$ , where  $y_i = \min(T_i, C_i)$ , and  $\delta = (\delta_1, \dots, \delta_n)'$ , where  $\delta_i = I_{[0, \infty)}(C_i - T_i) \equiv I(T_i \leq C_i)$ . The censoring times are assumed to be noninformative and to be independent of event times. In addition, we now have  $x_i$ , an  $r$  vector of predictor information associated with individual  $i$ .

### 13.1 Accelerated Failure Time Models

The AFT model is a log-linear regression model for event times  $T_1, \dots, T_n$ . The AFT model is

$$\log(T_i) = x_i' \beta + \sigma \varepsilon_i, \quad \varepsilon_i \stackrel{iid}{\sim} F_\varepsilon(\cdot) \quad (1)$$

where  $\varepsilon_i$  is an error term,  $\sigma = 1/\sqrt{\tau}$  is a scale parameter, and  $F_\varepsilon$  is a known cdf defined on the real line with corresponding density  $f_\varepsilon$ . We can also define “survival” and “hazard” functions  $S_\varepsilon(u) = 1 - F_\varepsilon(u)$  and  $h_\varepsilon(u) = f_\varepsilon(u)/S_\varepsilon(u)$ , although these are not restricted to nonnegative values  $u$ . When model (1) holds, write

$$T_i \stackrel{ind}{\sim} AFT(F_\varepsilon, \beta, \tau | x_i).$$

When  $F_\varepsilon = \Phi$ , the cdf of the  $N(0, 1)$ , model (1) reduces to the linear regression model of Chapter 9 applied to log-transformed event times. Of course, Chapter 9 did not deal with censored data. Other distributions used for  $F_\varepsilon$  include the logistic and extreme value (complementary log-log), cf. Table 13.1. These are the same three distributions used in Section 8.1 to define binomial regression models. We typically include an intercept parameter in  $x_i' \beta$  by setting  $x_{i1} \equiv 1$  and standardize any measurement predictor variables. Defining  $e_1 = (1, 0, \dots, 0)'$ , for standardized variables  $\beta_1 = e_1' \beta$  is the regression parameter for a “standard” individual in the study.

We need the survivor function for an individual with event time  $T$  and covariate information  $x$ , that is, the survival function when  $T \sim AFT(F_\varepsilon, \beta, \tau | x)$ . Model (1) implies

$$\begin{aligned} S(t | x, \beta, \tau) &= \Pr(T > t | x, \beta, \tau) \\ &= \Pr[\log(T) > \log(t) | x, \beta, \tau] \\ &= \Pr[(\log(T) - x' \beta) \sqrt{\tau} > (\log(t) - x' \beta) \sqrt{\tau} | x, \beta, \tau] \\ &= \Pr[\varepsilon > (\log(t) - x' \beta) \sqrt{\tau} | x, \beta, \tau] \\ &= S_\varepsilon[(\log(t) - x' \beta) \sqrt{\tau}] \\ &= 1 - F_\varepsilon[(\log(t) - x' \beta) \sqrt{\tau}]. \end{aligned} \quad (2)$$

The corresponding density and hazard functions derived from this are:

$$f(t | x, \beta, \tau) = \frac{\sqrt{\tau}}{t} f_\varepsilon[(\log(t) - x' \beta) \sqrt{\tau}] \quad (3)$$

$$h(t | x, \beta, \tau) = \frac{\sqrt{\tau}}{t} h_\varepsilon[(\log(t) - x' \beta) \sqrt{\tau}]. \quad (4)$$

Table 13.1: Common error distributions for AFT models.

Baseline Distribution	$f_\varepsilon(u)$	$F_\varepsilon(u)$	$W \equiv e^\varepsilon$
Normal	$\frac{1}{\sqrt{2\pi}}e^{-u^2/2}$	$\Phi(u)$	Log-normal
Logistic	$e^u/(1+e^u)^2$	$e^u/(1+e^u)$	Log-logistic
Extreme Value	$e^u \exp[-e^u]$	$1 - \exp[-e^u]$	Weibull

These functions depend on  $F_\varepsilon$ . If  $\varepsilon$  is standard normal,

$$f(t|x, \beta, \tau) = \frac{1}{\sqrt{2\pi}} \frac{\sqrt{\tau}}{t} \exp\left\{-\frac{\tau}{2}[\log(t) - x'\beta]^2\right\},$$

so similar to Example 12.2.3, we see that  $T$  is log-normal,  $T \sim LN(x'\beta, 1/\tau)$ . Using the extreme value cdf for  $F_\varepsilon$ , it is not too difficult to show that  $T \sim Weib(\sqrt{\tau}, e^{-x'\beta}\sqrt{\tau})$ , which is an exponential distribution whenever  $\tau = 1$ . To obtain this result, substitute the extreme value cdf for  $F_\varepsilon$  into (2) and simplify to recognize the Weibull survival function from Example 12.2.2.

The likelihood function based on a sample is constructed by substituting (2) and (3) into (12.1.2). Of course as indicated by Exercises 4.8 and 12.2(b), we could just as well use the distribution of  $\log(T)$ .

We have previously noted that the median of the  $LN(\mu, 1/\tau)$  distribution is  $e^\mu$ . The same argument gives the median of a  $LN(x'\beta, 1/\tau)$  distribution as  $e^{x'\beta}$ . In an arbitrary AFT model, let  $\tilde{t}_\varepsilon$  satisfy

$$0.5 = F_\varepsilon(\tilde{t}_\varepsilon).$$

For  $T \sim AFT(F_\varepsilon, \beta, \tau|x)$ , the median time to event  $\tilde{t}$  satisfies

$$0.5 = S(\tilde{t}|x, \beta, \tau).$$

From (2),

$$0.5 = 1 - F_\varepsilon[(\log(\tilde{t}) - x'\beta)\sqrt{\tau}],$$

so we must have

$$\tilde{t}_\varepsilon = (\log(\tilde{t}) - x'\beta)\sqrt{\tau}$$

or

$$\tilde{t} = \exp(x'\beta + \tilde{t}_\varepsilon/\sqrt{\tau}).$$

We are often interested in how  $\tilde{t}$  varies with  $x$ , so write  $\tilde{t}(x)$ .

If  $\tilde{t}_\varepsilon = 0$ , we get  $\tilde{t}(x) = \exp(x'\beta)$ . The standard normal and logistic distributions both have  $\tilde{t}_\varepsilon = 0$ . Often the extreme value distribution is redefined so that it also has  $\tilde{t}_\varepsilon = 0$ . This is accomplished by redefining the extreme value cdf as

$$F_\varepsilon(u) \equiv 1 - \exp[-\log(2)e^u].$$

The relative median comparing an individual with covariate  $x_1$  to another individual with covariate  $x_2$  is always

$$RM \equiv e^{(x_1 - x_2)'\beta}.$$

Often we are interested in estimating survival probabilities (2), and hazard functions (4) for selected choices of  $x$ , as well as hazard ratios

$$HR = \frac{h(t|x_1, \beta, \tau)}{h(t|x_2, \beta, \tau)} = \frac{h_\varepsilon[(\log(t) - x_1'\beta)\sqrt{\tau}]}{h_\varepsilon[(\log(t) - x_2'\beta)\sqrt{\tau}]}.$$

**EXAMPLE 13.1.1.** *Cancer of the Larynx: Log-Normal Regression.* Data on 90 males with cancer of the larynx were collected by Kardaun (1983) and analyzed in Klein and Moeschberger (2003). The data consist of times in months from diagnosis to death or censoring. Predictor variables include the Stage of the disease at diagnosis (1–4), the year of diagnosis ( $Yr$ ), and the age at diagnosis (Age). The Stage variable is recoded into indicators  $S_i$  for Stages 2, 3, and 4, making Stage 1 the baseline stage. We also standardize Age and  $Yr$  before using them. Our log-normal AFT model is

$$\log(T) = \beta_1 + \beta_2 S_2 + \beta_3 S_3 + \beta_4 S_4 + \beta_5 \text{Age} + \beta_6 Yr + \sigma \varepsilon,$$

with  $\varepsilon \sim N(0, 1)$ .

For people with the same age and year of diagnosis, we are interested in the median lifetimes for individuals in Stages 2, 3 or 4 relative to Stage 1. These are  $e^{\beta_i}$ ,  $i = 2, 3, 4$ . The relative median that compares two individuals who are one (sample) standard deviation (10.8 years) apart in age, but with the same stage and year of diagnosis, is  $e^{\beta_5}$ . The relative median that compares two individuals of the same stage and age but who differ by one standard deviation (2.2 years) in year of diagnosis is  $e^{\beta_6}$ .

We handle right censored observations in WinBUGS as in Subsection 12.2.5. With  $NA$  used to indicate missing data, transform the data  $(y, \delta)$  into  $(t, c)$  where

$$t_i = \begin{cases} y_i & \text{if } \delta_i = 1 \\ NA & \text{if } \delta_i = 0 \end{cases} \quad c_i = \begin{cases} 0 & \text{if } \delta_i = 1 \\ y_i & \text{if } \delta_i = 0 \end{cases}.$$

The WinBUGS code below for the log-normal regression model using reference priors includes a subset of the data listed at the end.

```
model{
  for(i in 1:106){
    sAge[i] <- (age[i]-mean(age[ ]))/sd(age[ ])
    sYr[i] <- (Yr[i]-mean(Yr[ ]))/sd(Yr[ ])
    t[i] ~ dlnorm(mu[i], tau) I(c[i],)
    mu[i] <- beta[1] + beta[2]*equals(stage[i],2)
                  + beta[3]*equals(stage[i],3)
                  + beta[4]*equals(stage[i],4)
                  + beta[5]*sAge[i] + beta[6]*sYr[i]
    # 5 month survival probabilities using covariates in the data
    S[i] <- 1- phi((log(5)-mu[i])*sqrt(tau))
    # Medians corresponding to the covariates in the data
    med[i] <- exp(mu[i])
  }
  for(i in 1:6){
    beta[i] ~ dnorm(0,0.000001)
    rm[i] <- exp(beta[i]) # RMs for each variable
    prob[i] <- step(beta[i])
  }
  tau ~ dgamma(0.001,0.001)
  sigma <- sqrt(1/tau)
}
list(beta=c(0,0,0,0,0,0),tau=1) # Initial values
stage[ ] t[ ] age[ ] Yr[ ] c[ ]
1 0.6 77 76 0
1 1.3 53 71 0
1 2.4 45 71 0
1 NA 57 78 2.5
1 3.2 58 74 0
```

```

1 NA 51 77 3.2
The last 84 data lines have been removed to save space
1 NA 50 71 0 # 16 lines of augmented data start here
1 NA 50 77 0
1 NA 70 71 0
1 NA 70 77 0
2 NA 50 71 0
2 NA 50 77 0
2 NA 70 71 0
2 NA 70 77 0
3 NA 50 71 0
3 NA 50 77 0
3 NA 70 71 0
3 NA 70 77 0
4 NA 50 71 0
4 NA 50 77 0
4 NA 70 71 0
4 NA 70 77 0
END

```

The last 16 lines in the data list contain no actual data since the time of death is *NA* and the censoring time is 0 in each line. These define predictor vectors (i.e., types of people) for which we wish to make inferences. We picked age, year, and stage values for the predictors that allow us to assess the practical importance of changing these variates on the median time to death and 5-month survival prospects. We could also compute predictive densities for the time to death of 16 new larynx cancer patients with these predictor vectors.

Posterior results on regression coefficients, relative survival medians, and the error scale factor  $\sigma$  are presented in Table 13.2. There is very little evidence that Stage 2 leads to shorter survival than Stage 1, but there is substantial evidence that Stages 3 and 4 are progressively worse than Stage 1. (Compare  $\beta_j$  to 0 or  $e^{\beta_j}$  to 1,  $j = 2, 3, 4$ .) We are 95% sure that the median time to death under Stage 1 is between  $3 \pm 1/0.36$  and  $20 = 1/0.05$  times longer than it is under Stage 4. Based on posterior medians, there is some indication that older patients have worse survival prospects (estimated  $\beta_5 < 0$ ) and that treatments are improving over time since later year of diagnosis has estimated  $\beta_6 > 0$ . The 95% posterior probability intervals for  $\beta_5$  and  $\beta_6$  both contain 0, so the evidence for these indications is not overwhelming. More directly we have  $\Pr(\beta_5 < 0 | D) = 0.90$  and  $\Pr(\beta_6 > 0 | D) = 0.74$ . The latter probability indicates a non-null effect of age at diagnosis on survival prospects.

If the probability intervals for  $\beta_5$  and  $\beta_6$  both contained 0 and were very narrow, we would remove *Age* and *Yr* from the analysis. However, Table 13.3 indicates that these variables have somewhat interesting effects on inferences. The table provides information on median survival times for 50 and 70 year-olds who were diagnosed in 1971 or in 1977. For example, Stage 2 individuals of age 50 and diagnosed in 1971 have an estimated median time to death of about 9 months, while if they were diagnosed in 1977, their estimated time to death is about 12 months. The 97.5 percentiles differ by nearly one year. Comparing 70 year-olds, these numbers are about 6 and 8 months, respectively. Interestingly these values are lower but have the same relative change. *Age* and *Yr* display interesting but modest effects that, especially with *Yr*, we cannot be sure are real. When considering the comparable Stage 4 results, the effects of *Yr* and *Age* are so small that scientifically they seem practically meaningless, although perhaps less so to larynx cancer patients. The only unambiguous stage effects are that Stages 3 and 4 give worse survival prospects than Stage 1, and that Stage 4 is worse than Stage 3, since  $\Pr(e^{\beta_3 - \beta_4} > 1 | Y) = 0.981$ , which implies that we are 98% sure that median survival in Stage 3 is longer than in Stage 4. The point estimates indicate 3 times longer. Finally, 5-month survival probabilities seem better for 50 year-olds diagnosed as Stage 1 in 1977

Table 13.2: Log-normal regression fit to the larynx cancer data.

Node	mean	sd	2.5%	median	97.5%
$\beta_1$	2.33	0.32	1.75	2.32	3.01
$\beta_2$	-0.25	0.51	-1.25	-0.25	0.76
$\beta_3$	-0.97	0.41	-1.81	-0.96	-0.17
$\beta_4$	-2.02	0.53	-3.09	-2.00	-1.01
$\beta_5$	-0.22	0.17	-0.56	-0.22	0.10
$\beta_6$	0.13	0.19	-0.24	0.12	0.52
$e^{\beta_2}$ St2/St1	0.89	0.49	0.29	0.78	2.11
$e^{\beta_3}$ St3/St1	0.41	0.17	0.17	0.38	0.83
$e^{\beta_4}$ St4/St1	0.15	0.08	0.05	0.14	0.36
$e^{\beta_3-\beta_4}$ St3/St4	3.23	1.83	1.07	2.80	7.89
$e^{\beta_5}$ (Age +1sd)/Age	0.81	0.14	0.57	0.80	1.11
$e^{\beta_6}$ (Yr +1sd)/Yr	1.15	0.22	0.80	1.12	1.63
$\sigma$	1.40	0.17	1.12	1.38	1.76

Table 13.3: Median survival times and 5-month survival probabilities estimated from a log-normal regression analysis of the larynx data.

	Node	mean	sd	2.5%	median	97.5%
Stage 2 Medians	Age 50 Yr 71	10.6	6.9	3.1	8.9	28.2
	Age 50 Yr 77	14.5	9.9	4.5	12.0	39.3
	Age 70 Yr 71	6.8	3.8	2.3	5.9	16.3
	Age 70 Yr 77	9.3	5.1	3.4	8.0	22.5
Stage 4 Medians	Age 50 Yr 71	1.83	1.18	0.48	1.55	4.93
	Age 50 Yr 77	2.42	1.40	0.80	2.10	5.99
	Age 70 Yr 71	1.18	0.66	0.36	1.03	2.83
	Age 70 Yr 77	1.55	0.72	0.62	1.40	3.35
Stage 1 5 Month Surv.	Age 50 Yr 71	0.72	0.10	0.51	0.73	0.88
	Age 50 Yr 77	0.78	0.10	0.56	0.79	0.94
	Age 70 Yr 71	0.62	0.09	0.43	0.62	0.79
	Age 70 Yr 77	0.69	0.10	0.48	0.70	0.87
Stage 3 5 Month Surv.	Age 50 Yr 71	0.46	0.12	0.24	0.46	0.70
	Age 50 Yr 77	0.55	0.12	0.31	0.55	0.78
	Age 70 Yr 71	0.35	0.10	0.17	0.35	0.56
	Age 70 Yr 77	0.44	0.11	0.23	0.43	0.66

than for 70 year-olds diagnosed in 1971 as Stage 3 because the posterior probability intervals do not overlap.

We noted earlier that  $T \sim AFT(F_\epsilon, \beta, \tau|x)$  with the extreme value cdf for  $F_\epsilon$  leads to  $T \sim \text{Weib}(\sqrt{\tau}, e^{-x'\beta}\sqrt{\tau})$ . To simplify interpretations of regression coefficients, we use the modified extreme value distribution that has median 0, i.e.,

$$F_\epsilon(u) = 1 - \exp[-\log(2)e^u],$$

which leads to

$$T \sim \text{Weib}(\sqrt{\tau}, \log(2)e^{-x'\beta}\sqrt{\tau}).$$

Table 13.4: Weibull regression fit to the larynx cancer data.

Node	mean	sd	2.5%	med	97.5%
$\beta_1$	2.22	0.29	1.71	2.20	2.86
$\beta_2$	-0.14	0.49	-1.07	-0.15	0.86
$\beta_3$	-0.66	0.38	-1.45	-0.65	0.05
$\beta_4$	-1.69	0.46	-2.64	-1.68	-0.83
$\beta_5$	-0.21	0.16	-0.54	-0.21	0.09
$\beta_6$	0.09	0.18	-0.24	0.08	0.45
$e^{\beta_2}$ St2/St1	0.99	0.54	0.35	0.87	2.37
$e^{\beta_3}$ St3/St1	0.55	0.21	0.24	0.52	1.04
$e^{\beta_4}$ St4/St1	0.20	0.10	0.07	0.19	0.43
$e^{\beta_5}$ (Age +1sd)/Age	0.82	0.13	0.58	0.81	1.109
$e^{\beta_6}$ (Year +1sd)/Year	1.1	0.2	0.78	1.08	1.58
$\sigma$	1.007	0.14	0.77	0.99	1.32

Table 13.5: Median survival times and 5-month survival probabilities estimated from a Weibull regression analysis of the larynx data.

	Node	mean	sd	2.5%	med	97.5%
Stage 2	Age 50 Yr 77	13.64	8.67	4.86	11.42	35.72
Medians	Age 70 Yr 77	8.67	5.26	3.43	7.71	22.3
Stage 4	Age 50 Yr 77	2.75	1.36	1.10	2.45	6.19
Medians	Age 70 Yr 77	1.77	0.66	0.87	1.7	3.39
Stage 1	Age 50 Yr 77	0.76	0.09	0.55	0.77	0.90
5 Month Surv.	Age 70 Yr 77	0.67	0.10	0.45	0.68	0.84
Stage 3	Age 50 Yr 77	0.59	0.12	0.33	0.60	0.81
5 Month Surv.	Age 70 Yr 77	0.47	0.13	0.22	0.47	0.71

We now use this form of Weibull regression to reanalyze the larynx cancer data.

EXAMPLE 13.1.2. *Cancer of the Larynx: Weibull Regression.* Using the same reference priors as in the log-normal analysis, the main change in the WinBUGS code given earlier is that the line

```
t[i] ~ dlnorm(mu[i], tau) I(c[i],)
```

is replaced by the code

```
t[i] ~ dweib(alpha,lambda[i]) I(c[i],)
lambda[i] <- log(2)*exp(-mu[i]*sqrt(tau))
```

and we define alpha by including the line

```
alpha <- sqrt(tau)
```

at the end of the program (or elsewhere but not inside a “for” loop). In addition, replace the log-normal 5-month survival probability

```
S[i] <- 1 - phi((log(5)-mu[i])*sqrt(tau))
```

with the Weibull 5-month survival probability

```
S[i] <- exp(-log(2)*exp((log(5)-mu[i])*sqrt(tau)))
```

Posterior results appear in Tables 13.4 and 13.5. Comparing these results with Tables 13.2 and 13.3 we see that the two analyses are similar. Using the log-normal model, medians comparing 50 year-olds to 70 year-olds diagnosed in 1977 were estimated as 12 and 8 in Stage 2, and 2.1 and

1.4 in Stage 4. Under the Weibull model these estimates are 11.4 and 7.7 in Stage 2, and 2.45 and 1.7 in Stage 4. Estimated relative medians comparing Stages 3 and 4 to Stage 1 under the *LN* were 0.38 and 0.14, while they are 0.52 and 0.19 under the Weibull. For the Weibull regression, the posterior probability that  $\beta_5$  is negative is 0.91 and the posterior probability that  $\beta_6$  is positive is 0.68. These probabilities were 0.90 and 0.74 under the *LN* model. The DIC statistics are 297.3 and 297.7 for *LN* and Weibull regression, respectively. The distribution of  $\sigma$  is tightly packed around one, which indicates that an exponential model may suffice. *If we are convinced that both models are reasonable, we should not draw any conclusions that are not supported by both.*

**EXERCISE 13.1.** Fit an exponential regression model to the larynx cancer data. This is accomplished by modifying the Weibull code so that  $\sigma$  is identically one. Obtain the DIC statistic and use it to decide if the Exponential model is adequate. Compare inferences with those in Tables 13.4 and 13.5.

Before reanalyzing the data with  $F_\varepsilon$  being the logistic distribution, we discuss this distribution in more detail. Suppose  $\varepsilon$  has the standard logistic distribution of Table 13.1. Write

$$\varepsilon \sim \text{Logis}(0, 1).$$

For positive  $\sigma$  and real  $\mu$  consider

$$W = \mu + \sigma\varepsilon.$$

Write the distribution of  $W$  as

$$W \sim \text{Logis}(\mu, 1/\sigma) = \text{Logis}(\mu, \tau^*)$$

where

$$\tau^* \equiv 1/\sigma.$$

Using arguments similar to those in (2), it is easy to see that  $W$  has cdf

$$F_W(w) = F_\varepsilon[(w - \mu)\tau^*] = 1 - \frac{1}{1 + \exp[(w - \mu)\tau^*]}.$$

If  $\log(T) \sim \text{Logis}(\mu, \tau^*)$ , define  $T$  to have a *log-logistic distribution* and write

$$T \sim LL(\mu, \tau^*).$$

**EXAMPLE 13.1.3. Cancer of the Larynx: Log-logistic Regression.** Now consider  $T_i \stackrel{\text{ind}}{\sim} AFT(F_\varepsilon, \beta, \tau | x_i)$  with the logistic error distribution of Table 13.1. Using (2),

$$S(t | x, \beta, \tau) = 1 - F_\varepsilon[(\log(t) - x'\beta)\sqrt{\tau}] = \frac{1}{1 + \exp\{(\log(t) - x'\beta)\sqrt{\tau}\}}$$

and

$$T_i \stackrel{\text{ind}}{\sim} LL(x_i'\beta, \sqrt{\tau}).$$

Equivalently,

$$\log(T_i) \stackrel{\text{ind}}{\sim} \text{Logis}(x_i'\beta, \sqrt{\tau}).$$

We use the same reference priors as in the log-normal analysis. The main change to the WinBUGS code is that the analysis is performed on the log scale, using the logistic distribution rather than the log-logistic. As usual in WinBUGS, transform  $(y, \delta)$  into  $(t, c)$  but now use  $(\log(t), \log(c))$  in the WinBUGS analysis. When  $t$  or  $c$  is zero, replace it with a large negative number, here,  $-1000$ . When  $t = \text{NA}$ , define  $\log(t) = \text{NA}$ . As discussed before, if  $t$  and  $c$  remain in the data file, they must be used in the WinBUGS program.

```

junk1 <- t[1]
junk2 <- c[1]
stage[ ] t[ ] logt[ ] age[ ] Yr[ ] c[ ] logc[ ]
1      0.6   -0.5108  77     76     0    -1000
1      1.3    0.2623  53     71     0    -1000
1      2.4    0.8754  45     71     0    -1000
1      NA     NA      57     78     2.5   0.9162
1      3.2    1.1631  58     74     0    -1000
1      NA     NA      51     77     3.2   1.1631
More data lines and augmented data
END

```

The other key change from the log-normal code is that

```
t[i] ~ dlnorm(mu[i], tau) I(c[i],)
```

is replaced by

```

logt[i] ~ dlogis(mu[i], taustar)I(logc[i],)
taustar <- sqrt(tau)

```

As with alpha in the Weibull example, taustar must be assigned outside the “for” loop or an error will occur in WinBUGS. In addition, the log-normal 5-month survival probability

```
S[i] <- 1 - phi((log(5)-mu[i])*sqrt(tau))
```

is replaced with the log-logistic 5-month survival probability

```
S[i] <- 1/(1 + exp((log(5)-mu[i])*sqrt(tau)))
```

To create the variables logt and logc we copied the original data from WinBUGS into Excel. There we divided it into separate columns using the “Text to Columns” tool located under the Data menu in the Excel toolbar. Then we used the `ln` function in Excel to create a new column of log transformed survival times (if  $t_i$  is missing then set its log to NA) and a column of log transformed censoring times (if  $c_i = 0$  then set the log equal to  $-1000$  or some other large negative value). Finally, we copied that data from Excel back into WinBUGS. In WinBUGS go to Paste Special under the Edit menu and select Unicode Text. As always, you need to type END with a hard return after the last data line. Then you are set to run the code with the transformed data.

The results of the log-logistic analysis look superficially similar to those of the log-normal and Weibull. We leave as an exercise the analysis of the larynx cancer data using log-logistic regression and comparing it to the other models.

Unfortunately, comparisons between the three models are complicated by computational differences in the analyses. The log-logistic analysis is actually a logistic analysis of log transformed data whereas the first two were log-normal and Weibull analyses of the data rather than normal and extreme value analyses of log transformed data. As established in Exercises 4.8 and 12.2(b), the likelihood functions based on censored samples from original and transformed data are proportional to one another, so applications of Bayes’ Theorem are identical. However, equivalence of transformed and untransformed data does not extend to many of our model comparison criteria. For example, both the log-normal and Weibull models have DIC  $\approx 297$ . Naive application of DIC to the logistic model for log transformed data gives DIC = 231, which *looks much better but is not really comparable*. Recall the discussion at the beginning of Subsection 4.9.4 that the constant  $C$  differs when comparing say a log-normal model for data to a normal model for the log transformation of the same data. We now examine this issue in more detail.

With  $T \sim AFT(F_\varepsilon, \beta, \tau)$ , the density for an observation looks like (3), while if we define  $W = \log(T)$  and let  $(W - x'\beta)\sqrt{\tau} \sim F_\varepsilon$ , then the density for  $W$  is  $\sqrt{\tau}f_\varepsilon[(w - x'\beta)\sqrt{\tau}]$ , where  $w = \log(t)$ . The likelihood functions are proportional and the ratio of the likelihood function based on the AFT model (3) to that based on the transformed data is  $1/\prod_{i=1}^n y_i^{\delta_i}$ . Obviously, this is free of the parameters. When calculating a DIC in WinBUGS based on the log transformed data model,

the term  $C \equiv 2 \sum_{i=1}^n \delta_i \log(y_i)$  is missing relative to the DIC based on model (3). Note that  $p_D$  is the same regardless of which way the model is specified since the constant cancels in its calculation. To properly compare the DIC for our transformed data version of the log-logistic model to the DICs for the log-normal and Weibull models that both use form (3), we need to add  $C$  to the value 231 obtained earlier.

**EXERCISE 13.2.** (a) Analyze the larynx cancer data using the log-logistic regression model. Compare results with the previous analyses. You will need to calculate the value  $C$  defined above and add that to the value of DIC obtained for the logistic model based on log transformed data. (b) Pick one of the three models used to analyze these data, and investigate whether adding interaction terms improves the selected model. If it does, modify the model interpretation accordingly.

### 13.1.1 Abortion Data

In the remainder of this section, our discussion revolves around the data introduced in Section 1.4 on time to natural abortion in dairy cattle. We focus on inferential procedures using an informative prior. The next subsection details the elicitation of the informative prior. The last three subsections examine case deletion diagnostics, model selection, and sensitivity analysis.

The 45 data records were retrieved from a larger set of 375 observations collected by Drs. Mark Thurmond and Sharon Hietala of the University of California, Davis. While all 375 cows were at risk of abortion, only 45 actually aborted. Our analysis is conditional on the event of a known abortion. Most of the remaining cows had full-term pregnancies. Some cows were culled from the herd, but there is no way to tell if they would have aborted or gone full term. We make the assumption that culled animals would have come to term.

The data involve four covariates. In Section 1.4 we only used infection status to distinguish two groups. The variable  $x_2 = IS$  is an indicator of infection status wherein infected with *Neospora caninum* corresponds to  $x_2 = 1$  and no infection corresponds to zero. Days open,  $x_3 = DO$ , is the number of days between most recent previous birth and conception. Calving ease,  $x_4 = CE$ , is (inexplicably) coded  $x_4 = 1$  for cows with difficulties in their most recent previous birth and zero otherwise. Prostaglandin,  $x_5 = PR$ , has  $x_5 = 1$  corresponding to prostaglandin being given during the open days. Along with an intercept  $x_1 \equiv 1$ , the scientists believed there may be an interaction  $x_6 = DO * CE$ , so there are 6 regression coefficients in our full model:

$$\log(T_i) = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_6 x_{i6} + \sigma \varepsilon_i.$$

The scientists believed that infection status would be the most important variable for predicting time to abortion, with abortion occurring later in infected animals. They expected days open to be the second most important variable, anticipating that increasing  $DO$  would slightly increase the time to abortion. Difficulty with calving and the application of prostaglandin were expected to slightly decrease the time to abortion. The primary goal of this study is to characterize the effect of neospora on the time to abortion. This is part of an ongoing investigation of neospora and other infectious abortifacients with the ultimate goal of reducing fetal wastage, cf. Paré, Thurmond, and Hietala (1997).

Figure 13.1 presents eight predictive survival curves for various choices of  $x = (1, x_2, x_3, x_4, x_5, x_6)'$  using our elicited prior and a logistic error distribution. Also presented are median survival times for each set of covariates. These curves represent cows with  $PR = 0$ . We also constructed similar plots for  $PR = 1$ . Superimposing the plots shows no visual differences between them; however, predicted median survival times are a few days longer for  $PR = 0$  than for  $PR = 1$ . These differences are of little practical importance. For example, in plot (b), for  $x' = (1, 1, 45, 1, 0, 45)$ , the estimated median time to abortion is 114 days. The 90% PI is (92, 141), so clearly a couple of days difference in estimated median survival times is of no importance. Similarly, in plot (a), for  $x' = (1, 0, 45, 0, 0, 0)$ , the median time to abortion is 63 days with 90% PI of (54, 74). Note also that when  $CE = 1$ , the curves

depend on the infection status but do not seem to depend much on the days open. On the other hand, when  $CE = 0$ , fetal survival depends on both the infection status and the number of days open.

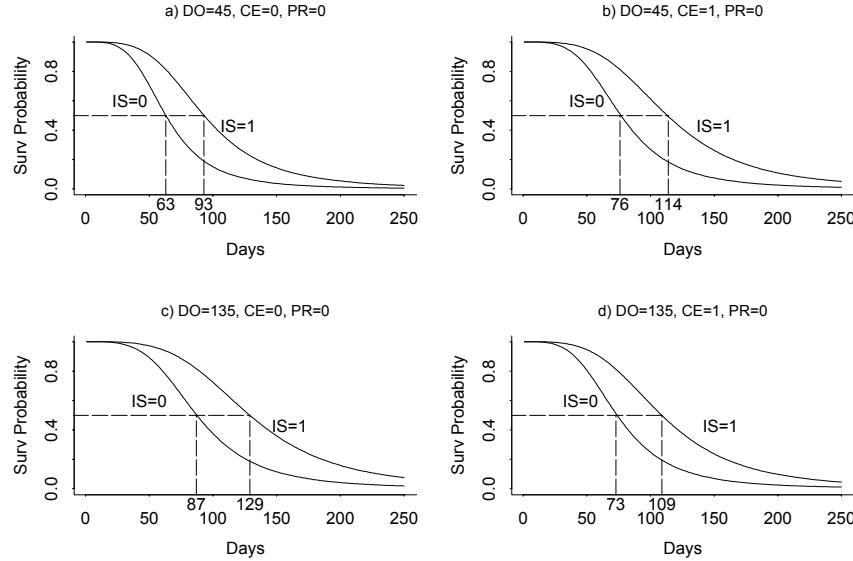


Figure 13.1: Predictive survival curves for cow abortion study.

We present summary values based on posterior and maximum likelihood methods in Table 13.6. The posterior means are similar to the maximum likelihood estimates. Standard deviations are either about the same or smaller for the posterior. The larger discrepancies are due to having little data on some conditions with significant prior information. Statistical import or lack thereof is clear from looking at the various probability intervals for all the factors. Everything but prostaglandin seems to matter as we have already seen from looking at the survivor curves.

Predictive survival probabilities can be thought of as the probability of survival for a particular individual with covariate  $x$ , or they can be thought of as an estimate of the proportion of individuals who survive in a population of individuals with covariate  $x$ . With the latter interpretation we can get approximate posterior distributions for  $h(\beta, \sigma) = S(t|x, \beta, \sigma) = 1 - F_\varepsilon((\log(t) - x'\beta)/\sigma)$  and, therefore, obtain posterior probability bands for these proportions, i.e., PIs for every  $t > 0$ .

### 13.1.2 Prior Elicitation for AFTs

We require a joint prior  $p(\beta, \tau)$ . The SIR prior for AFT models is  $p(\beta, \tau) \propto 1/\tau$ , cf. Ibrahim and Laud (1991). We used a proper approximation to this prior for all three analyses of the larynx cancer data. A common alternative to the SIR prior is the prior from Section 9.3, which is conjugate for uncensored normal or log-normal data, but which is not conjugate for censored data or for data that are not (transformed) normal. This prior lets  $\beta$  be normal given  $\tau$  and the marginal of  $\tau$  be Gamma, cf. West (1985) and Racine-Poon et al. (1986). Normal-gamma priors are also convenient for large samples, where the posterior distribution of  $\beta$  is approximately normal. However, repeating our mantra, it is difficult to specify real prior information directly for esoteric regression parameters.

Bedrick, Christensen, and Johnson (2000) used informative conditional median priors. Their prior specification is made on a collection of median responses, each corresponding to a potential

Table 13.6: Log-logistic regression estimates for the cow abortion study.

Variable	Informative Posterior Summaries			
	Mean	Std. Dev.	5%	95%
Intercept	3.985	0.146	3.742	4.223
<i>IS</i>	0.388	0.086	0.247	0.525
<i>DO</i>	0.004	0.001	0.002	0.006
<i>CE</i>	0.377	0.201	0.045	0.711
<i>PR</i>	-0.020	0.101	-0.182	0.146
<i>DO</i> * <i>CE</i>	-0.004	0.002	-0.007	-0.001
$\sigma$	0.261	0.032	0.213	0.314
Variable	Maximum Likelihood		SIR	Prior
	Estimate	S.E.	Mean	Std. Dev.
Intercept	3.928	0.152	3.929	0.160
<i>IS</i>	0.342	0.128	0.340	0.134
<i>DO</i>	0.004	0.001	0.004	0.001
<i>CE</i>	0.349	0.902	0.294	1.028
<i>PR</i>	-0.019	0.289	-0.035	0.315
<i>DO</i> * <i>CE</i>	-0.004	0.006	-0.004	0.007
$\sigma$	0.245	0.031	0.269	0.033

covariate combination. The resulting joint prior induces a distribution on the regression coefficients. Our informative priors assume  $\beta \perp\!\!\!\perp \tau$ , so we discuss the marginals for  $\beta$  and  $\tau$  separately.

### 13.1.2.1 Specifying the Marginal Prior for $\beta$

For AFT models with an error distribution that has median 0, the median survival time is  $m \equiv \exp(x'\beta)$ . We elicit expert information about a collection of medians and use that information to induce a prior on  $\beta$ . As we have done previously, select  $r$  linearly independent  $r \times 1$  covariate vectors,  $\tilde{x}_i = (1, \tilde{x}_{i2}, \tilde{x}_{i3}, \dots, \tilde{x}_{ir})'$ . Define the corresponding median times to event  $\tilde{m}_i = \exp(\tilde{x}'_i \beta)$ , and elicit expert opinion about these quantities. The induced distribution on  $\beta$  is obtained from the prior distribution on  $\tilde{m} \equiv (\tilde{m}_1, \dots, \tilde{m}_r)'$ . Defining  $\tilde{X} = (\tilde{x}_1, \dots, \tilde{x}_r)'$ , write the vector  $\tilde{m} \equiv \exp(\tilde{X}\beta)$ . Since  $\tilde{X}$  is nonsingular, we obtain  $\beta = \tilde{X}^{-1} \log(\tilde{m})$ . The  $\tilde{x}_i$ s must be “far enough apart” that it is reasonable to assume that knowledge about the  $\tilde{m}_i$ s can be regarded independently. If  $\tilde{p}_i(\tilde{m}_i)$  is our prior on  $\tilde{m}_i$ , then  $\tilde{p}(\tilde{m}) = \prod_{i=1}^r \tilde{p}_i(\tilde{m}_i)$  is our prior on  $\tilde{m}$  and by Proposition B.4

$$p(\beta) \propto \prod_{i=1}^r \tilde{p}_i(e^{\tilde{x}'_i \beta}) e^{\tilde{x}'_i \beta},$$

where  $\prod_{i=1}^r e^{\tilde{x}'_i \beta}$  is proportional to the Jacobian of the transformation. To see this, observe

$$\frac{d}{d\beta'} \exp(\tilde{X}\beta) = \begin{bmatrix} \frac{d}{d\beta'} \exp(\tilde{x}'_1 \beta) \\ \vdots \\ \frac{d}{d\beta'} \exp(\tilde{x}'_r \beta) \end{bmatrix} = \begin{bmatrix} \tilde{x}'_1 \exp(\tilde{x}'_1 \beta) \\ \vdots \\ \tilde{x}'_r \exp(\tilde{x}'_r \beta) \end{bmatrix} = \text{Diag}\{e^{\tilde{x}'_i \beta}\} \tilde{X}$$

thus

$$\left| \det \left[ \frac{d}{d\beta'} \exp(\tilde{X}\beta) \right] \right| = \left| \det \left[ \text{Diag}\{e^{\tilde{x}'_i \beta}\} \tilde{X} \right] \right| = |\det(\tilde{X})| \prod_{i=1}^r e^{\tilde{x}'_i \beta} \propto \prod_{i=1}^r e^{\tilde{x}'_i \beta}.$$

Because  $\tilde{m}$  is on the scale of the data, it is more natural to think about than  $\beta$ . Moreover, the prior on  $\tilde{m}$  induces the same prior on  $\beta$  for any median zero error distribution  $F_\epsilon$ . Thus a single elicitation allows us to evaluate various AFT models.

Since median survival times are non-negative, a convenient prior has

$$\tilde{m}_i \stackrel{\text{ind}}{\sim} LN(\log(\tilde{m}_{0i}), \tilde{\sigma}_{0i}^2) \quad (5)$$

or

$$\log(\tilde{m}_i) \stackrel{\text{ind}}{\sim} N(\log(\tilde{m}_{0i}), \tilde{\sigma}_{0i}^2).$$

The expert's best guess for  $\tilde{m}_i$  is taken as the median of the prior,  $\tilde{m}_{0i}$ . The prior standard deviation  $\tilde{\sigma}_{0i}$  is obtained by eliciting information on an upper (or lower) percentile of  $\tilde{m}_i$ , exactly as in Subsections 5.2.3 and 9.4.1.

WinBUGS code for inducing a prior on  $\beta$  based on independent log-normal priors for the  $\tilde{m}_i$ s is presented below. The inverse function in WinBUGS is designed to handle symmetric positive definite matrices so we always compute  $\tilde{X}^{-1}$  outside of WinBUGS.

```
model{
  for(i in 1:r){
    mtilde[i] ~ dlnorm(mu0[i],tautilde0[i]) # mtilde[i] ~ LN
    mu0[i] <- log(mtilde0[i]) # Prior guess is mtilde0[i]
    tautilde0[i] <- 1/pow(sigmatilde0[i],2)
    lmtilde[i] <- log(mtilde[i])
    beta[i] <- inprod(Xtildeinv[i,1:r], lmtilde[1:r])
  }
}
list(r =, mtilde0 = c(), sigmatilde0 = c())
# The matrix Xtildeinv can be obtained in R
# using Xtildeinv <- solve(Xtilde)
Xtildeinv[,1] Xtildeinv[,2] Xtildeinv[,3] ... Xtildeinv[,r]
list the columns of Xtildeinv here
END
```

**EXERCISE 13.3. Log-normal Prior.** (a) Let  $r = 2$ . Obtain explicit log-normal priors for  $\tilde{m}_1$  and  $\tilde{m}_2$  if the best guesses for the two medians are 10 and 20, and if the values of the 95th percentiles are 20 and 30, respectively. (b) Assume that the single predictor variable is age and that it has been standardized, that the average age in the data is 50, and that the standard deviation is 5. Write code to induce the prior on  $\beta$  when you have selected an average age of 50 for the first predictor and an age of 60, two standard deviations above the average, for the second. Then

$$\tilde{X} = \begin{pmatrix} 1 & 0 \\ 1 & 2 \end{pmatrix}.$$

(c) Re-write the code for the log-normal so that you solve two equations in two unknowns. Run the code and show the induced priors on the  $\beta_i$ s.

Sometimes, as with the abortion data, it is useful to have more flexibility in modeling prior information than just the log-normal priors of (5). The normal priors on  $\log(\tilde{m}_i)$  can be generalized to priors based on the other median 0 error distributions  $F_\epsilon$ . Choose a prior of the form

$$\frac{\log(\tilde{m}_i) - \log(\tilde{m}_{0i})}{\sigma_{0i}} \stackrel{\text{ind}}{\sim} F_\epsilon. \quad (6)$$

The prior (5) is a special case of (6) with  $F_\epsilon$  a standard normal cdf. Bedrick, Christensen, and Johnson (2000) considered log-extreme-value (Weibull) distributions as priors for regression coefficients in the Weibull regression model, and log-logistic distributions as priors for regression coefficients in the log-logistic regression model. They also used a combination of log-normal and Weibull priors for the cow abortion data to match elicited prior information better.

Similar to the computation for normal errors, the prior density for  $\beta$  is obtained from the prior (6) on  $\tilde{m}$  via  $\beta = \tilde{X}^{-1} \log(\tilde{m})$  with the Jacobian of the transformation remaining the same. We will sample the prior on  $\tilde{m}$  in WinBUGS, so we don't need the analytical form of the prior density for  $\beta$ .

To identify a prior of the form (6), select  $\tilde{m}_{0i}$  exactly as before. Let  $w_\alpha$  be the  $1 - \alpha$  percentile of the distribution  $F_\epsilon$ . If  $\Pr(\tilde{m}_i < c) = 1 - \alpha$ , we have

$$w_\alpha = \frac{\log(c) - \log(\tilde{m}_{0i})}{\sigma_{0i}}$$

and solve for  $\tilde{\sigma}_{0i}$ .

If we have the median 0 extreme value error distribution, the percentile  $w_\alpha$  is easily found. For example, if  $\alpha = 0.05$ , then  $0.95 = F_\epsilon(w_{0.05}) = 1 - \exp[-\log(2)e^{w_{0.05}}]$ , so  $\log(0.95) = -\log(2)e^{w_{0.05}}$  and

$$w_{0.05} = \log\{[\log(20)/\log(2)]\} = 1.46.$$

WinBUGS code for inducing a prior on  $\beta$  based on an *LL* specification for the  $\tilde{m}_i$ s is:

```
model{
  for(i in 1:r){
    w[i] ~ dlogis(mu0[i],tautilde0[i])
    mtilde[i] <- exp(w[i]) # So mtilde[i] ~ Log-Logistic
    mu0[i] <- log(mtilde0[i])
    tautilde0[i] <- 1/sigmatilde0[i]
    lmtilde[i] <- log(mtilde[i])
    beta[i] <- inprod(Xtildeinv[i,1:r], lmtilde[1:r])
  }
}
```

WinBUGS code for inducing a prior on  $\beta$  based on a Weibull specification for the  $\tilde{m}_i$ s is:

```
model{
  for(i in 1:r){
    mtilde[i] ~ dweib(alphatilde0[i],thetatilde0[i])
    thetatilde0[i] <- log(2)*exp(-mu0[i]*sqrt(tautilde0[i]))
    alphatilde0[i] <- sqrt(tautilde0[i])
    mu0[i] <- log(mtilde0[i])
    tautilde0[i] <- 1/pow(sigmatilde0[i],2)
    lmtilde[i] <- log(mtilde[i])
    beta[i] <- inprod(Xtildeinv[i,1:r], lmtilde[1:r])
  }
}
```

**EXERCISE 13.4. *BCJ Priors.*** Repeat parts (a) and (b) of Exercise 13.3 for the logistic and for the extreme value error distributions.

**EXAMPLE 13.1.4. *Cow Abortion Data.*** We elicited prior information from Dr. Mark Thurmond, a veterinary epidemiologist. We ascertained his beliefs about the median time to abortion at six sets of covariates. He specified best guesses for the median time to abortion (the medians of his prior distributions) and extreme values for his distributions, that we took to be the 5th and 95th percentiles. The covariate combinations were chosen within the range of Dr. Thurmond's experience and far enough apart so that median times to abortion for the various covariate combinations could be regarded independently. The six covariate combinations are linearly independent so that  $\tilde{X}^{-1}$  exists.

The information is in Table 13.7. The rows correspond to “locations,” with  $i = 1$  corresponding to uninfected cows ( $IS = 0$ ), with 95 days open ( $DO = 95$ ), no prior history of difficult births

Table 13.7: Specification of the prior for cow abortions.

$i$	Int.	Design for Prior					Prior Percentiles			$\tilde{\sigma}_{0i}$
		IS	DO	CE	PR	DO * CE	5%	$\tilde{m}_{0i}$	95%	
1	1	0	95	0	0	0	70	80	95	0.1045
2	1	0	45	1	1	45	60	75	90	0.1108
3	1	0	45	1	0	45	65	75	90	0.1108
4	1	1	130	1	1	130	80	110	130	0.1141
5	1	1	130	0	1	0	80	120	140	0.1053
6	1	1	95	0	0	0	85	125	150	0.1246

( $CE = 0$ ), and no prostaglandin given ( $PR = 0$ ). Dr. Thurmond's best guess for the median time to abortion was 80 days with only a 5% chance that the median time to abortion would be less than 70 days and a 95% chance that the median time to abortion would be less than 95 days. Similarly,  $i = 4$  corresponds to infected cows with 130 days open, a history of difficult births, and prostaglandin administered; his best guess for the median time to abortion was 110 days with only a 5% chance that the median time to abortion would be less than 80 days and a 95% chance that the median time to abortion would be less than 130 days. Note the strong influence of infection on the prior median times to abortion. Dr. Thurmond tended to be more certain about situations with smaller median times as one might expect, e.g., percentiles tend to be closer when the median is smaller.

Dr. Thurmond was generally more confident about the 95th percentiles than the 5th percentiles, so we determined Weibull and log-normal distributions that matched the medians and 95th percentiles listed in the table. Log-normal distributions were used for cases 1, 2, and 3 and Weibulls were used for the others. The choice between Weibull and log-normal was based on how well their 5th percentiles matched the elicited 5th percentiles (the log-normal being, for these values, more right skewed and the Weibull more left skewed). The distributions were plotted and we reconfirmed that they appropriately represented the expert's prior opinions.

Figure 13.2 presents plots of the prior distributions for the  $\tilde{m}_i$ s and the corresponding posteriors for the median times to abortion under each of the six covariate combinations  $\tilde{x}_i$ . Except at the very center, the prior and posterior are nearly identical for location 2. This is because there are only a few animals among the 45 that were given prostaglandin. There is little information in the data for this location and the prior information dominates. At the other locations, there is generally more information in the posterior than the prior. The priors and posteriors for locations 2, 3, and 5 have approximately the same centers while the centers for the posteriors at locations 1, 4, and 6 are less than for the corresponding priors.

### 13.1.2.2 Partial Prior Information for $\beta$

As discussed in Subsections 8.4.4 and 9.4.3, we may not have the resources to specify a full prior for  $r$  medians. Our approach here is similar to the earlier discussions. After standardizing all continuous covariates, we specify independent prior distributions on, say,  $p < r$  medians,  $\tilde{m}_i = e^{\tilde{x}_i' \beta}, i = 1, \dots, p$  corresponding to a  $p \times r$  matrix  $\tilde{X} = (\tilde{X}_1, 0)$ , where  $\tilde{X}_1$  is  $p \times p$  and nonsingular. Partition  $\beta = (\beta_1^*, \beta_2^*)'$ . Then the prior has been specified on the vector of medians  $\tilde{m}_1^* = \exp(\tilde{X}_1 \beta_1^*)$ . Solving for  $\beta_1^* = \tilde{X}_1^{-1} \log(\tilde{m}_1^*)$ , the prior on  $\beta_1^*$  is induced by the specification for  $\tilde{m}_1^*$ . We then specify an independent reference prior for the components of  $\beta_2^*$ , say, independent  $N(0, 1000)$  priors.

**EXERCISE 13.5.** (a) For Exercise 13.3 write WinBUGS code to induce a partial prior on  $\beta$  where  $p = 1, r = 2$  using the first row of  $\tilde{X}$  to identify  $\tilde{m}_1^* \equiv \tilde{m}_1$ . Repeat using the second row of  $\tilde{X}$  and let  $\tilde{m}_1^* \equiv \tilde{m}_2$ . (b) Now suppose that there is a second covariate, say a dichotomous variate indicating

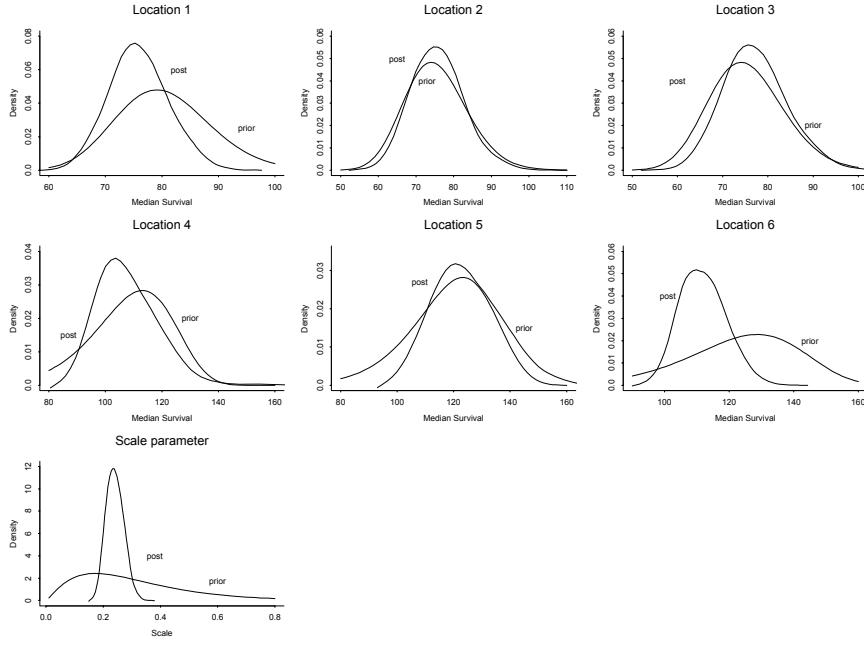


Figure 13.2: Prior and posterior densities for  $\tilde{m}_i$ s and for  $\sigma$  in the cow abortion study.

Sex ( $M = 1, F = 0$ ). Let

$$\tilde{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 2 & 0 \end{pmatrix}$$

and assume that the prior information for the two medians in Exercise 13.3 was actually specified for women. Using this partial prior information, write code to induce a prior on  $\beta$ .

#### 13.1.2.3 Uncertainty About $\tau$

The prior on  $\tau$  is elicited as in Example 5.2.2, which involved log transformed data from the Diasorin study. For an AFT model with error median 0, consider a percentile of the survival distribution for  $T$  other than the median. Information about this percentile gives us information about  $\tau$  and  $\sigma = 1/\sqrt{\tau}$ . We focus on  $\sigma$ ; the prior for  $\tau$  is similar.

Consider the 25th percentile of the survival distribution for cows with covariate vector  $\tilde{x}_1$ . Suppose the experimenter's best guess for the median time to event for such cows is  $\tilde{m}_{01} = 80$  and his best guess for the 25th percentile is 60. With  $F_\varepsilon(w_{0.25}) = 0.25$ , the 25th percentile of  $T$  is  $\exp(\tilde{x}_1\beta + \sigma w_{0.25}) = \tilde{m}_1 \exp(\sigma w_{0.25})$ . The experimenter has asserted that the median of his distribution for  $\tilde{m}_1 \exp(\sigma w_{0.25})$  was 60, given that  $\tilde{m}_1 = 80 = \tilde{m}_{01}$ . Using the fact that  $\tilde{m}$  and  $\sigma$  are independent (since we assume  $\beta$  and  $\tau$  are independent),

$$\begin{aligned} 0.5 &= \Pr[\tilde{m}_1 \exp(\sigma w_{0.25}) \leq 60 \mid \tilde{m}_1 = 80] \\ &= \Pr\left[\frac{\log[\tilde{m}_1 e^{(\sigma w_{0.25})}] - \log[\tilde{m}_1]}{w_{0.25}} \geq \frac{\log(60/80)}{w_{0.25}} \mid \tilde{m}_1 = 80\right] \\ &= \Pr\left[\sigma \geq \frac{-0.2877}{w_{0.25}}\right], \end{aligned}$$

where the direction of the inequality has changed since  $w_{0.25} < 0$ . The meaning of  $\sigma$  depends on the error distribution used to model the data, so to go further we need to specify the error distribution. For logistic errors,  $w_{0.25} = -1.0986$ , thus

$$0.5 = \Pr[\sigma \geq 0.262]$$

and 0.262 is the median of the prior distribution for  $\sigma$ .

Dr. Thurmond was 95% certain that the value of the quartile of  $T$  was at least 37, leading to another percentile of the distribution of  $\sigma$ . These percentiles approximately match a  $\text{Gamma}(2.25, 7.25)$  distribution for  $\sigma$ , which is plotted with the posterior in the bottom row of Figure 13.2.

### 13.1.3 Case Deletion Diagnostics for AFT Models

Johnson (1996) introduced case deletion diagnostics for estimating survival curves in the Bayesian log-normal survival model. As discussed in Section 9.6, the goal is to detect individual cases that noticeably affect inferences. For example, it is important to know whether the predictive survivor function changes radically upon deleting one case. It is also relevant to examine the effect of the prior on posterior inferences and, with a BCJ prior, it is possible to look at the effect of deleting a “prior observation” corresponding to  $\tilde{m}_i$ .

First we give an expression for the joint posterior with a case deleted. The complete data are  $D \equiv (y, \delta)$ . Denote the data with the  $i$ th case deleted by  $D_{(i)}$ . Write the  $i$ th case data as  $d_i = (y_i, \delta_i)$ . The likelihood based on all the data except  $d_i$  is  $L(\beta, \sigma | D_{(i)}) \equiv L(\beta, \sigma | D) / L(\beta, \sigma | d_i)$ , so

$$\begin{aligned} p(\beta, \sigma | D_{(i)}) &= \frac{L(\beta, \sigma | D_{(i)}) p(\beta, \sigma)}{\int L(\beta, \sigma | D_{(i)}) p(\beta, \sigma) d\beta d\sigma} \\ &= \frac{p(\beta, \sigma | D) / L(\beta, \sigma | d_i)}{\int p(\beta, \sigma | D) / L(\beta, \sigma | d_i) d\beta d\sigma}. \end{aligned} \quad (7)$$

**EXERCISE 13.6.** Prove equation (7).

With a sample from the full data joint posterior, say  $\{(\beta^k, \sigma^k) : k = 1, \dots, m\}$ , it is possible to make inferences about functions of the parameters with a case deleted without obtaining another posterior sample. Suppose we are interested in making inferences about  $\gamma(\beta, \sigma)$  using the case deleted posterior. From (7),

$$\begin{aligned} E[\gamma(\beta, \sigma) | D_{(i)}] &= \int \gamma(\beta, \sigma) p(\beta, \sigma | D_{(i)}) d\beta d\sigma \\ &= \frac{\int \{\gamma(\beta, \sigma) / L(\beta, \sigma | d_i)\} p(\beta, \sigma | D) d\beta d\sigma}{\int \{1 / L(\beta, \sigma | d_i)\} p(\beta, \sigma | D) d\beta d\sigma} \\ &\doteq \frac{\sum_k \gamma(\beta^k, \sigma^k) / L(\beta^k, \sigma^k | d_i)}{\sum_k 1 / L(\beta^k, \sigma^k | d_i)}. \end{aligned}$$

To obtain an approximation to the case deleted predictive density  $f(t | x, D_{(i)})$  at a particular value  $t$ , take

$$\gamma(\beta, \sigma) = f(t | x, \beta, \sigma) = \frac{1}{\sigma t} f_\epsilon \{(\log(t) - x' \beta) / \sigma\}.$$

In practice, this is obtained on a grid of equally spaced points  $\{t_1, t_2, \dots, t_s\}$ . While connecting the dots provided by the pairs  $[t_k, f(t_k | x, D_{(i)})]$  gives a reasonable approximation to the predictive density, we can facilitate computation of predictive probabilities and predictive expectations by

creating a discrete approximation to the continuous predictive distribution. Define the approximate discrete density as

$$\hat{f}(t|x, D_{(i)}) \equiv \frac{f(t|x, D_{(i)})}{\sum_{j=1}^s f(t_j|x, D_{(i)})}, \quad t \in \{t_1, \dots, t_s\},$$

with a similar definition of  $\hat{f}(t|x, D)$ . Define the corresponding approximations to the predictive survivor functions to be  $\hat{S}(t|x, D)$  and  $\hat{S}(t|x, D_{(i)})$ , which are obtained by summing (numerically integrating) the functions  $\hat{f}(t|x, D)$  and  $\hat{f}(t|x, D_{(i)})$  over the grid points.

A standard tool for measuring the effect of case deletion on prediction (estimation) is the Kullback-Leibler divergence (KLD) between predictive (posterior) densities based on full and case deleted data (cf. Johnson and Geisser, 1983, 1985; Johnson, 1996). The KLD is a measure of the discrepancy between two densities. If the predictive density based on the full data is appreciably different from the predictive density based on case deletion, we should take a closer look at the case in question. See also the similar discussion in Section 9.6 for linear regression. We introduce a variation of the approach of Johnson (1996) for the log-normal survival model.

Consider first the effect of case deletion on the joint posterior. Define the KLD diagnostic

$$K_i = \int p(\beta, \sigma|D) \log \left( \frac{p(\beta, \sigma|D)}{p(\beta, \sigma|D_{(i)})} \right) d\beta d\sigma.$$

This simplifies using (7) and some algebra to

$$\begin{aligned} K_i &= \log \left( \int \frac{1}{L(\beta, \sigma|d_i)} p(\beta, \sigma|D) d\beta d\sigma \right) - \\ &\quad \int \log \left( \frac{1}{L(\beta, \sigma|d_i)} \right) p(\beta, \sigma|D) d\beta d\sigma \\ &= \log \left( E \left[ \frac{1}{L(\beta, \sigma|d_i)} \right] \right) - E \left[ \log \left( \frac{1}{L(\beta, \sigma|d_i)} \right) \right] \geq 0 \end{aligned}$$

where both expectations are with respect to the full posterior and nonnegativity is guaranteed by Jensen's inequality. A numerical approximation is

$$K_i \doteq \log \left( \frac{1}{m} \sum_{k=1}^m \left[ 1/L(\beta^k, \sigma^k|d_i) \right] \right) - \frac{1}{m} \sum_{k=1}^m \log \left[ 1/L(\beta^k, \sigma^k|d_i) \right].$$

Compute this for all cases, censored and uncensored. Recall that

$$L(\beta, \sigma|d_i) = \frac{1}{\sigma y_i} f_\varepsilon[(\log(y_i) - x'_i \beta)/\sigma]^{\delta_i} \{1 - F_\varepsilon[(\log(y_i) - x'_i \beta)/\sigma]\}^{1-\delta_i},$$

where  $d_i = (y_i, \delta_i)$  is the observed data for case  $i$ .

There is no natural scale for  $K_i$  so we normalize values by dividing by the largest. Cases with unusually large values should be scrutinized as discussed in Section 9.6. *If there is a mistake associated with a case, it can either be fixed or the case can be removed from the data. Otherwise, the model should be run with unusual cases deleted and a determination made whether those cases cause any substantial impact on inferences. If not, proceed with the analysis based on the full data. If there is a substantial impact, report how inferences change if unusual cases are deleted.*

**EXERCISE 13.7.** For log-normal regression of the larynx cancer data, take the Monte Carlo output from WinBUGS into R and obtain the values of  $K_i$  for all subjects in the study. Find and remove the most influential case and re-run the analysis. Comment on the impact of deleting that case on all inferences.

### 13.1.3.1 Predictive Influence

We develop two measures of the effect of case deletion on prediction. The first is a KLD for assessing the effect of case deletion on a predictive density corresponding to a predictor  $x$ . This is defined as

$$KP_i^x = \int f(t|x,D) \log \left( \frac{f(t|x,D)}{f(t|x,D_{(i)})} \right) dt.$$

We also propose a measure that directly examines the effect of case deletion on the right tail of survival curves. Ferguson (1973) employed  $\int [F(t) - \hat{F}(t)]^2 f(t) dt$  as a loss function when using  $\hat{F}$  to estimate a cdf  $F$ . To emphasize the right tails of the survival function, we transform to logs and consider a similar measure,

$$DS_i^x = \int f(t|x,D) \{ \log[S(t|x,D)/S(t|x,D_{(i)})] \}^2 dt.$$

$DS_i^x$  will be large if case  $i$  has a substantial effect on the right tail of the predictive density. The two measures can be approximated by

$$KP_i^x \doteq \sum_{j=1}^s \hat{f}(t_j|x,D) \log \left( \frac{\hat{f}(t_j|x,D)}{\hat{f}(t_j|x,D_{(i)})} \right),$$

and

$$DS_i^x \doteq \sum_{j=1}^s \hat{f}(t_j|x,D) \left\{ \log \left( \frac{\hat{S}(t_j|x,D)}{\hat{S}(t_j|x,D_{(i)})} \right) \right\}^2.$$

We can examine the collective effect of deleting case  $i$  over a future sample of individuals with covariates, say,  $x_{f\ell}$ ,  $\ell = 1, \dots, L$ , using

$$KP_i = \sum_{\ell=1}^L KP_i^{x_{f\ell}}, \quad DS_i = \sum_{\ell=1}^L DS_i^{x_{f\ell}}.$$

Johnson and Geisser (1983) let  $L = n$  and  $x_{f\ell} = x_\ell$ , the predictor vectors associated with the observed data. Since there is no natural scale for  $KP_i$  or  $DS_i$ , we normalize all values by dividing by the largest.

**EXAMPLE 13.1.5. Cow Abortion Data.** We considered the eight locations used in Figure 13.1 as our choices of  $x_{f\ell}$ 's. Case 10 was the most influential with normalized values of 1 on both predictive measures. This was followed by case 26 with  $KP_{26} = 0.96$  and  $DS_{26} = 0.82$ . The third most influential value based on  $KP$  was case 2 with  $KP_2 = 0.47$ , but for this case  $DS_2 = 0.00$ , indicating that case 2 had some impact on the predictive densities but virtually no impact on the tails of the corresponding survival functions. Finally, case 38 was third most influential with respect to the  $DS$  measure with  $DS_{38} = 0.76$ , while  $KP_{38} = 0.14$ . Thus, case 38 had a relatively large impact on estimation of the tails of the survival distribution but had little impact on the predictive densities.

We note that cases 10, 26, and 38 had by far the highest observed times to abortion: 259, 254, and 251 days, respectively. These pregnancies nearly went full term (about 260 days). The next highest survival time was 150 days.

**EXERCISE 13.8.** Reanalyze the cow abortion data without cases 10, 26, and 38 and compare the results to the full data analysis.

### 13.1.4 Bayes Factor Model Selection

Which error distribution is most appropriate for AFT data? Equivalently, how can we select from among the *LL* ( $M_1$ ), *LN* ( $M_2$ ) and Weibull ( $M_3$ ) regression models? We have illustrated using DIC for this task. Additionally, the LPML statistic can be used to construct pseudo Bayes factors as it was for linear regression. We now consider calculating Bayes factors as discussed in Section 4.8 and Subsection 8.3.2. Sections 4.8 and 4.9 contain basic concepts of all of these methods and Sections 8.3 and 9.7 contain applications. The same basic techniques apply here.

In earlier applications of Bayes factors with uncensored data, we found the marginal density of  $y$  from model  $M$ . However, now the data are  $D = (y, \delta)$  because they are subject to censoring. Naively, we might write the marginal density for the data as

$$f(D|M) = \int f(D|\theta, M)p(\theta|M)d\theta,$$

where the definition of  $\theta$  depends on the model. Moreover, we might consider simulating from the marginal distribution of  $D$  by using the method of composition. This usually entails simulating  $\theta^k \sim p(\theta|M)$  and then simulating  $y^k \sim f(y|\theta^k, M)$ . However, without knowledge of the censoring mechanism, we cannot hope to simulate censored data. Recalling our derivation in Subsection 12.1.3 of the likelihood function for censored data under the assumptions of noninformative censoring and independence of censoring and failure times, we know that  $f(D|\theta, M) = h(D)L(\theta|D, M)$  for some function  $h(\cdot)$  that depends on the unknown censoring distribution. Thus,

$$f(D|M) = h(D) \int L(\theta|D, M)p(\theta|M)d\theta.$$

Although  $h(\cdot)$  can never be known, in computing Bayes factors it cancels, so we can simply ignore  $h(\cdot)$ .

As a function of an AFT model  $M$ ,  $f(D|M)$  is proportional to integrating the likelihood function with respect to the prior on  $(\beta, \sigma)$  for the model. The Bayes factor for comparing models  $M_j$  and  $M_{j'}$  is

$$BF_{jj'} = \frac{f(D|M_j)}{f(D|M_{j'})} = \frac{\int L(\beta, \sigma|D, M_j)p(\beta, \sigma|M_j)d\beta d\sigma}{\int L(\beta, \sigma|D, M_{j'})p(\beta, \sigma|M_{j'})d\beta d\sigma}.$$

**EXAMPLE 13.1.6. Cow Abortion Data.** The Bayes factors for the 3 pairwise comparisons of AFT models under the BCJ prior are  $BF_{12} = 2.6$ ,  $BF_{13} = 4922$ , and thus (with some round-off error)  $BF_{23} = BF_{13}/BF_{12} = 4922/2.6 = 1880$ . There seems to be good reason not to use Weibull regression. Log-logistic regression appears to be the best among these candidates, but the log-normal does not seem too bad. In particular, if the prior odds for the log-logistic versus log-normal models are 1, the posterior odds in favor of the log-logistic are 2.6.

### 13.1.5 Sensitivity Analysis

We redid the analysis of the cow abortion data using the SIR prior  $p(\beta, \tau) \propto 1/\tau$ . (This analysis was not performed using WinBUGS.) Survival curves similar to Figure 13.1 are given in Figure 13.3. Again, curves for  $PR = 1$  are visually almost indistinguishable from those shown for  $PR = 0$ . The curves for  $IS = 0$ ,  $CE = 0$  are almost indistinguishable from those using our informative prior. The curves for  $IS = 1$ ,  $CE = 0$  are slightly lower than with our prior, indicating slightly shorter survivals. Predicted median survivals dropped by just under 10 days. Posterior 90% intervals for these are about 30 days long, so a change of 10 days is not drastic.

Comparing plots 13.1(b) and 13.3(b), we see that the tails are much higher with the SIR prior and the curves are closer together, indicating less difference due to infection status. The SIR prior's posterior median survival of 98 days is nearly outside the informative prior's posterior 90% interval

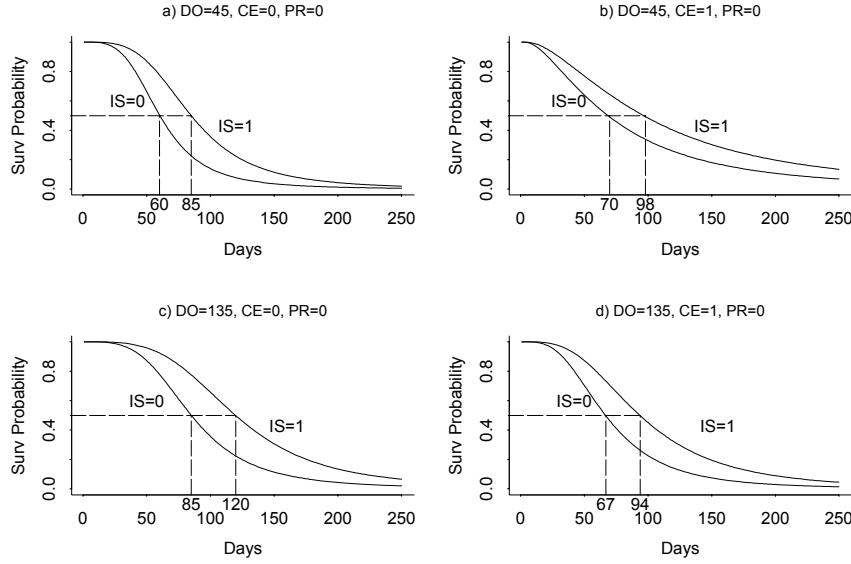


Figure 13.3: *Survival curve estimates for cow abortion study. SIR prior.*

of (92,141). What is really going on is that there is little information in the data corresponding to these conditions, so the informative prior is exerting considerable weight. As mentioned, the 90% informative posterior interval for  $IS = 1, DO = 45, CE = 1, PR = 0$  is (92, 141). For the SIR analysis, it is (27, 311). For  $IS = 0$  and the SIR prior, the interval is (19, 227). These wide intervals are indicative of little information from the data.

The differences between predictive fetal survival curves from the SIR and informative priors in plots 13.1(d) and 13.3(d) are larger than those for plots (a) and (c), but not nearly as extreme as those for the plots (b). Plots (a) and (c) have  $CE = 0$  and in the data there are many observations with  $CE = 0$ , so the prior has relatively little effect on these plots. There are only three observations with  $CE = 1$ , while the informative prior has three  $\hat{m}_i$ s with  $CE = 1$ , hence the larger differences between Figures 13.1 and 13.3 in plots (b) and (d).

Posterior means and standard deviations for the regression coefficients using the SIR prior were reported in Table 13.6. They are very similar to the corresponding maximum likelihood results that were also reported there.

#### 13.1.6 Final Comments

In Section 7.3 we referred to the use of generalized linear models to analyze time to event data. Rather than the AFT model (13.1.1), a generalized linear model based on the gamma distribution assumes

$$T_i \stackrel{ind}{\sim} \text{Gamma}(\alpha, \lambda_i)$$

so that

$$\text{E}(T_i) \equiv m_i = \alpha/\lambda_i.$$

Regression modeling in this context commonly takes the form

$$m_i = e^{x'_i \beta}, \quad \text{i.e.,} \quad \log(m_i) = x'_i \beta.$$

This *Gamma regression* model is not an AFT model except in the special case  $\alpha \equiv 1$  when  $\text{Gamma}(1, \lambda_i) = \text{Exp}(\lambda_i) = \text{Weib}(1, \lambda_i)$ .

All of the models in this section were based on distributions that depend on  $r + 1$  parameters. In the next section we consider an approach that allows more general distributional forms.

### 13.2 Proportional Hazards Modeling

In Section 13.1 we considered parametric regression models for time to event data. The application of these parametric regression models is somewhat restrictive. The distributions are all unimodal, which precludes the existence of clusters of individuals who survive longer than others. Such clusters often result in multimodal or exceptionally fat tailed survival distributions. In Section 15.1 we consider some models that could be used for the AFT error distribution that provide much more flexible density shapes. As an alternative, we now consider a broad family of models that also allows great flexibility, namely the Cox (1972) *proportional hazards* (PH) models. Specifically, we develop a version of the PH model that is amenable to WinBUGS analysis and illustrate its use with examples. Consult Ibrahim, Chen, and Sinha (2001) for developments in Bayesian proportional hazards modeling that go beyond ours.

#### 13.2.1 The Proportional Hazards (PH) Model

Proportional hazards regression seeks to flexibly model survival as a function of predictor information  $x$ . This is accomplished through modeling the hazard function as

$$h(t | x, \beta) = e^{x' \beta} h_0(t), \quad (1)$$

where  $h_0(\cdot)$  is a completely arbitrary hazard function that determines a *baseline distribution* with density  $f_0$ , cdf  $F_0$ , and survival curve  $S_0$ . Under model (1), as  $x' \beta$  becomes larger, the hazard of an event gets larger.

Unlike most regression models including AFT models, PH regression does not include an intercept. More properly, the vector  $x$  in the PH model is not assumed to have  $x_1 \equiv 1$ . An intercept would get confounded with the baseline hazard function  $h_0$ , cf. Exercise 13.10(a).

In our analysis, we will need the concept of the *cumulative hazard function*. The cumulative hazard is

$$H(t) \equiv \int_0^t h(s) ds.$$

We now show that

$$S(t) = \exp[-H(t)].$$

From (12.1.1),  $h(t) = f(t)/S(t)$ . Given this relation, it is not difficult to show that

$$h(t) = -\frac{d}{dt} \log[S(t)].$$

By the Fundamental Theorem of Calculus,

$$\log[S(t)] = - \int_0^t h(s) ds + c$$

for some constant  $c$ . This occurs if and only if

$$S(t) = \exp \left[ - \int_0^t h(s) ds + c \right].$$

Since the survivor function at  $t = 0$  must be 1, we must have  $c = 0$ , which completes the proof. In addition, since for large  $t$  the survival function  $S(t)$  approaches zero,  $H(t)$  must approach infinity, cf. Exercise 13.10(b).

Using model (1),

$$H(t|x,\beta) = \int_0^t h(s|x,\beta)ds = e^{x'\beta} \int_0^t h_0(s)ds = e^{x'\beta} H_0(t),$$

where  $H_0(\cdot)$  is the baseline cumulative hazard function. It follows that

$$S(t|x,\beta) = \exp[-e^{x'\beta} H_0(t)] = \{\exp[-H_0(t)]\}^{e^{x'\beta}} = [S_0(t)]^{e^{x'\beta}}. \quad (2)$$

The density function is

$$f(t|x,\beta) = h(t|x,\beta)S(t|x,\beta) = e^{x'\beta} h_0(t)[S_0(t)]^{e^{x'\beta}}.$$

If  $x'_1\beta > x'_2\beta$ , from (2) we have  $S(t|x_1,\beta) < S(t|x_2,\beta)$  for all  $t$ , since a probability to a greater power must be less than one to a smaller power. In the PH model, survival curves for individuals with distinct covariates never cross. The PH model is therefore inappropriate when covariates interact with time. For example, suppose that one of two treatments for cancer is more toxic early in a study so that survival prospects are worse under that treatment in the beginning. However, individuals who survive the toxicity may do better with this treatment later on, leading to crossing survival curves. Fortunately, there are many situations where such interactions between time and covariates (e.g., treatment) are not an issue and the PH model is appropriate. Moreover, when an interaction exists, it can often be remedied through the use of *time dependent covariates*. We discuss these near the end of the section (before the string of exercises) and in Section 15.3.

A key feature of the PH model is that for two distinct covariate vectors  $x_1$  and  $x_2$ , the hazard ratio (*HR*) is

$$HR = \frac{h(t|x_1,\beta)}{h(t|x_2,\beta)} = \exp[(x_1 - x_2)'\beta],$$

which does not depend on the time  $t$ . The hazard function in the numerator is equal to this constant *HR* times the hazard in the denominator, i.e.,

$$h(t|x_1,\beta) = HR \times h(t|x_2,\beta),$$

hence the name “proportional hazards model.”

While the hazard ratio is a simple function of  $\beta$  in model (1), median times to event and relative medians are not simple. We focus on survival curve estimates and relative hazards in PH models.

Consider standard survival data  $\{(y_i, \delta_i)\}_{i=1}^n$  arising from the PH model (1). Using (12.1.3) the likelihood function is

$$L(\beta, h_0) = \prod_{i=1}^n \{e^{x'_i\beta} h_0(y_i)\}^{\delta_i} [S_0(y_i)]^{e^{x'_i\beta}}. \quad (3)$$

**EXERCISE 13.9.** (a) Plot the hazard ratio comparing the  $LN(0, 1)$  to the  $LN(2, 2)$ . Are the hazards proportional? (b) Repeat part (a) for  $LL(0, 1)$  versus  $LL(2, 2)$  distributions. (c) Repeat for  $LEV(0,1)$  versus  $LEV(2,2)$  distributions where  $LEV$  indicates the log of the extreme value (Weibull) distribution and the parameters are location and scale parameters for the extreme value distribution.

**EXERCISE 13.10.** (a) Show that model (1) is equivalent to picking a value  $t_0$  and requiring  $h_0(t_0) = 1$  but allowing an intercept in the model. (b) Show that  $\lim_{t \rightarrow \infty} H_0(t) = \infty$ .

### 13.2.2 A Baseline Hazard Model

A key feature of the likelihood (3) is that  $h_0$  is an entire function. We could model  $h_0$  as a parametric hazard function, say the Weibull of Example 12.2.2 so that

$$h_0(t|\alpha, \lambda) = \lambda \alpha t^{\alpha-1},$$

which is a function of two parameters. Exercise 13.11 establishes that this PH model is equivalent to the AFT Weibull regression model.

**EXERCISE 13.11.** Show that the Weibull regression model, with an appropriate redefinition of  $\beta$ , satisfies the PH model (1). Give the explicit form for the baseline hazard  $h_0$ , the baseline cumulative hazard  $H_0$ , and the baseline survivor function  $S_0$ .

Unfortunately, there is a dearth of useful parametric models for hazard functions. Much of the popularity of the PH model is that Cox (1972) developed methods for estimating  $\beta$  without having to know the baseline hazard function. For a Bayesian analysis, we must have some model for the baseline hazard.

A simple yet flexible model for  $h_0$  is a step function. Pick  $a_k$ s with  $0 \equiv a_1 < a_2 < \dots < a_K < a_{K+1} = \infty$ . Define the piecewise constant function

$$h_0(t) = \sum_{k=1}^K \lambda_k I_{[a_k, a_{k+1})}(t). \quad (4)$$

This means that  $h_0(t) = \lambda_k$  for all  $t \in [a_k, a_{k+1})$ . The values  $\lambda_k$  are unknown parameters. Typically, we take  $\lambda_k > 0$  for all  $k$  because to have  $\lambda_k = 0$  implies  $\Pr[a_k \leq T < a_{k+1}] = 0$ . For  $h_0$  to be a valid hazard function, it must integrate to infinity (over its domain of definition, usually the positive real line). If it does not,  $S(t) = \exp[-H(t)]$  does not equal zero as time gets large. Having the step function (4) integrate to infinity requires  $\lambda_K > 0$ .

We've gone from an infinite dimensional, unspecified parameter function  $h_0$  to a  $K$  dimensional parameter vector  $\lambda = (\lambda_1, \dots, \lambda_K)'$ . If  $K$  is large, we have a flexible "richly parametric" model that approximates an arbitrary  $h_0$ . In our second illustration,  $K = 130$ . The likelihood is obtained by substituting (4) into (3) and writing

$$L(\beta, \lambda) = L(\beta, h_0).$$

In Section 15.3 we briefly discuss other nonparametric approaches to modeling the baseline hazard. The corresponding PH models are *semi-parametric* in that they include regression parameters for the covariates but a "nonparametric" model for the baseline hazard. (AFT models with "nonparametric" error distributions are also semi-parametric.) With any such approach, the key is in being able to specify a workable likelihood and prior. We now look at these issues in detail for the step function hazard model. Our discussion in Section 15.3 is more "cookbook" in that these key issues are not addressed specifically for the PH model.

### 13.2.3 The Likelihood

Unfortunately, WinBUGS does not know how to construct the likelihood (3) based on our hazard model (4). To generate the likelihood we use what has become a standard "trick" in survival analysis that relies on Poisson modeling of *expanded* or *reconstructed* data.

First, we rewrite the likelihood. In the likelihood (3), the  $i$ th term of the product is

$$\begin{aligned} L_i(\beta, h_0) &= \left\{ e^{x_i' \beta} h_0(y_i) \right\}^{\delta_i} \{S_0(y_i)\}^{e^{x_i' \beta}} \\ &= \left\{ e^{x_i' \beta} h_0(y_i) \right\}^{\delta_i} \{ \exp[-H_0(y_i)] \}^{e^{x_i' \beta}} \\ &= \left\{ e^{x_i' \beta} h_0(y_i) \right\}^{\delta_i} \exp \left\{ -e^{x_i' \beta} H_0(y_i) \right\}. \end{aligned}$$

For the hazard model (4), let  $i^*$  be the largest integer with  $a_{i^*} \leq y_i$ . Now

$$\begin{aligned} L_i(\beta, \lambda) &= \left\{ e^{x'_i \beta} \lambda_{i^*} \right\}^{\delta_i} \exp \left\{ -e^{x'_i \beta} \left[ \lambda_{i^*} (y_i - a_{i^*}) + \sum_{j=1}^{i^*-1} \lambda_j (a_{j+1} - a_j) \right] \right\} \\ &= \left\{ e^{x'_i \beta} \lambda_{i^*} \right\}^{\delta_i} \exp \left\{ -e^{x'_i \beta} \left( \sum_{j=1}^{i^*} H[i, j] \lambda_j \right) \right\}, \end{aligned} \quad (5)$$

where

$$H[i, j] \equiv \begin{cases} a_{j+1} - a_j & \text{if } y_i \geq a_{j+1} \\ y_i - a_j & \text{if } y_i \in [a_j, a_{j+1}) \\ 0 & \text{if } y_i < a_j \end{cases}.$$

In computing the likelihood,  $y_i$  is fixed, so for any  $j$ ,  $H[i, j]$  is a fixed constant that does not involve the parameters.

To execute the trick, begin by defining an indicator function that identifies the interval in which the  $i$ th individual “died” (if they died),

$$N[i, j] = \begin{cases} 1 & \text{if } y_i \in [a_j, a_{j+1}] \text{ and } \delta_i = 1 \\ 0 & \text{otherwise} \end{cases} \quad i = 1, \dots, n; j = 1, \dots, K.$$

In particular,  $N[i, i^*] = 1$  if  $i$  corresponds to a death time and  $N[i, j] = 0$  if  $j \neq i^*$  or if  $i$  is censored. Using (5), the contribution to the likelihood from the data for subject  $i$  is

$$L_i(\beta, \lambda) \propto \prod_{j=1}^{i^*} \left( e^{x'_i \beta} \lambda_j \right)^{N[i, j]} \exp \left( -e^{x'_i \beta} H[i, j] \lambda_j \right).$$

Define  $\Theta[i, j] = e^{x'_i \beta} H[i, j] \lambda_j$ . The trick is to rewrite the likelihood as a product of  $\text{Pois}(\Theta[i, j])$  densities. This involves replacing  $(e^{x'_i \beta} \lambda_j)^{N[i, j]}$  in the likelihood contribution with  $\{\Theta[i, j]\}^{N[i, j]}$ . We need to establish that the two terms are proportional, as functions of the parameters, for any  $H[i, j]$ . Trivially, the two terms are proportional whenever  $N[i, j] = 0$ . Moreover, since  $H[i, j]$  does not involve any of the parameters, if  $H[i, j] > 0$  the two terms are proportional (as functions of the parameters). To get the trick to work, we need the terms to be proportional even when  $H[i, j] = 0$ . In other words, we need  $H[i, j] = 0$  to imply that  $N[i, j] = 0$  so that when  $H[i, j] = 0$ , we have  $H[i, j]^{N[i, j]} = 1$  under the convention  $0^0 = 1$ . Since  $N[i, j] > 0$  only when  $j = i^*$  and  $i$  is a death, we need  $H[i, i^*] > 0$  whenever  $i$  is a death. This happens as long as *none of the death times occurs at an  $a_j$  value*. Given the sampling model, the probability of a death occurring at an  $a_j$  is zero, so this should not be a problem (unless we do something to make it a problem). The reason the trick fails when  $H[i, i^*] = 0$  for a death time is because the strictly positive term  $(e^{x'_{i^*} \beta} \lambda_{i^*})^{N[i, i^*]}$  is not proportional to 0 =  $\{\Theta[i, i^*]\}^{N[i, i^*]} = 0^1$  as a function of the parameters.

Substituting  $\Theta[i, j]$  into the likelihood contribution gives

$$L_i(\beta, \lambda) \propto \prod_{j=1}^{i^*} \{\Theta[i, j]\}^{N[i, j]} \exp\{-\Theta[i, j]\}. \quad (6)$$

This is the likelihood for  $N[i, j] \stackrel{\text{ind}}{\sim} \text{Pois}(\Theta[i, j])$ ,  $j = 1, \dots, i^*$  where the Poisson data all happen to be either 0 or 1. Multiplying the terms in (6) gives the full likelihood (3),

$$L(\beta, \lambda) = \prod_{i=1}^n L_i(\beta, \lambda) = \prod_{i=1}^n \prod_{j=1}^{i^*} \{\Theta[i, j]\}^{N[i, j]} \exp\{-\Theta[i, j]\}.$$

Observe that since  $N[i, j] = \Theta[i, j] = 0$  for  $j > i^*$ , the  $i^*$  in the second product can be replaced with  $K$ .

### 13.2.3.1 Noninformative Data\*

Together, (2) and (4) determine a rich parametric family of sampling distributions for the data. However, if the number of parameters ( $K + r$ ) exceeds the number of data points, the distribution will not be identifiable, cf. Section 4.14. Even more distressingly, it is easy for specific data to be uninformative about specific parameters.

It is not hard to see from (5) that if there are no observations in the last interval, the likelihood function does not involve  $\lambda_K$ . Let  $\lambda_{(K)}$  be the  $\lambda$  vector with the  $K$ th element removed. It follows from our discussion at the end of Section 4.14 on noninformative data that

$$\lambda_K | \lambda_{(K)}, D \sim \lambda_K | \lambda_{(K)}.$$

In particular, if  $\lambda_K$  and  $\lambda_{(K)}$  are independent in the prior (and they will be in our standard prior), the posterior distribution of  $\lambda_K$  is independent of  $\lambda_{(K)}$  and is exactly the same as its prior.

Another more complex instance of noninformative data occurs if there are two consecutive intervals, say,  $[a_j, a_{j+1})$  and  $[a_{j+1}, a_{j+2})$ , neither of which contain any observations. The likelihood depends on  $\lambda_j$  and  $\lambda_{j+1}$  only through  $\lambda_j(a_{j+1} - a_j) + \lambda_{j+1}(a_{j+2} - a_{j+1})$ . Thus a variety of  $\lambda_j$  and  $\lambda_{j+1}$  values give the same likelihood, so there is no specific information on either parameter. Technically, the conditional posterior and the prior are the same, i.e.,

$$\begin{aligned} \lambda_j, \lambda_{j+1} | \lambda_j(a_{j+1} - a_j) + \lambda_{j+1}(a_{j+2} - a_{j+1}), D \\ \sim \lambda_j, \lambda_{j+1} | \lambda_j(a_{j+1} - a_j) + \lambda_{j+1}(a_{j+2} - a_{j+1}). \end{aligned}$$

Alternatively, we could reparameterize  $\lambda$  into  $\xi$  where  $\lambda = \xi$  except that we replace the parameters  $\lambda_{j+1}$  and  $\lambda_j$  with  $\xi_{j+1} = \lambda_j(a_{j+1} - a_j) + \lambda_{j+1}(a_{j+2} - a_{j+1})$  and, say,  $\xi_j = \lambda_{j+1} - \lambda_j$ . Let  $\xi_{(j)}$  be the  $\xi$  vector with the  $j$ th element removed. Now the likelihood does not involve  $\xi_j$ , so there is no information in the data about  $\xi_j$  and

$$\xi_j | \xi_{(j)}, D \sim \xi_j | \xi_{(j)}.$$

In particular, if  $\xi_j$  and  $\xi_{(j)}$  were independent in the prior (something that is quite unlikely), the posterior distribution of  $\xi_j$  would be independent of  $\xi_{(j)}$  and exactly the same as its prior.

### 13.2.4 Priors for $\beta$

Throughout we assume

$$\beta_j \stackrel{\text{ind}}{\sim} N(\beta_{0j}, \sigma_{0j}^2).$$

We need to identify the  $\beta_{0j}$ s and  $\sigma_{0j}^2$ s. Reference priors use  $\beta_{0j} = 0$  and  $\sigma_{0j}^2$  large.

To obtain an informative prior for  $\beta$ , we need a scale that facilitates elicitation of expert opinion. Researchers familiar with survival methods commonly think about relative hazards, so we elicit prior information on hazard ratios. A hazard ratio of 1 indicates no effect. A hazard ratio of 100 or 1/100 is something to write home about. Often, an *HR* of 1.5 or 2/3 is hardly worth mentioning, but an *HR* of 2 or 1/2 is.

Consider two subjects who are identical except that they differ on covariate  $j$  by 1 unit. These might be two individuals with, say, one a year older than the other. Their hazard ratio is  $HR_j = e^{\beta_j}$ . If the expert's best guess is  $HR_{0j}$ , then  $\beta_{0j} = \log(HR_{0j})$ . If a best guess for the 95th percentile of  $HR_j$  is  $u_j$ ,

$$0.95 = \Pr[e^{\beta_j} \leq u_j] = \Phi\left(\frac{\log(u_j) - \log(HR_{0j})}{\sigma_{0j}}\right).$$

This implies that  $1.645 = [\log(u_j) - \log(HR_{0j})]/\sigma_{0j}$ , which occurs if and only if  $\sigma_{0j} = [\log(u_j) - \log(HR_{0j})]/1.645$ .

It may be easier to think about an *HR* that corresponds to a larger difference than 1 unit. For instance, we may ask the expert to think about individuals who are 10 years apart so that  $HR = e^{10\beta_j}$ .

Our best prior estimate is then  $HR_0 = e^{10\beta_{0j}}$ , so  $\beta_{0j} = \log(HR_0)/10$ . Similarly, letting  $u$  denote the 95th percentile for  $HR$ ,  $\sigma_{0j} = [\log(u)/10 - \log(HR_0)/10]/1.645$ . This is equivalent to saying that  $10\beta_j \sim N(\log(HR_0), \{[\log(u) - \log(HR_0)]/1.645\}^2)$ . Note that the PH model presupposes that the hazard ratio for a 10-year difference is the same regardless of whether the individuals are 20 and 30 or whether they are 60 and 70.

Specifying  $e^{\beta_j}$  is relatively easy for binary predictors. For example, a researcher may believe that the hazard of male drivers being involved in an alcohol-related accident is triple that for females, so  $\beta_{0i} = \log(3)$ .

With interaction terms in the model, we have to work a little harder because we may not be able to identify individuals who differ only in one predictor by one unit. Suppose our predictors are Sex ( $F = 1, M = 0$ ) and Race (Black = 1, White = 0). With an interaction term, model (1) becomes

$$h(t | \text{Sex, Race, Sex} \times \text{Race}, \beta) = \exp(\beta_1 \text{Sex} + \beta_2 \text{Race} + \beta_3 \text{Sex} \times \text{Race}) h_0(t).$$

The interaction variable is 1 if and only if Race and Sex are both 1. There is no way to fix Race and Sex while looking at different levels of interaction.

Our baseline individual here is a White Male (WM). His hazard function is simply  $h_0$ . Consider the three  $HR$ s corresponding to comparisons between (i) a WF and a WM, (ii) a BM and a WM, and (iii) a BF and a WM. The three hazard ratios are (i)  $HR_1 \equiv e^{\beta_1}$ , (ii)  $HR_2 \equiv e^{\beta_2}$ , and (iii)  $HR_I \equiv e^{\beta_1 + \beta_2 + \beta_3}$ . We ask our expert to ponder all three of these and to provide their best estimates of each and an upper or lower value. We illustrate with upper values. Let  $HR_{0j}$ s denote the best guesses and  $u_j$ s denote the upper 95% values. The first two are easy; we have  $\beta_{0j} = \log(HR_{0j})$ ,  $j = 1, 2$ . The corresponding variances are determined just as with no interaction. Without interaction in the model, comparing WF-WM is the same as comparing BF-BM, but with interaction we need to specify the Race and, based on our coding of the variables, to get the simple  $HR$  of  $e^{\beta_1}$  it must be specified as White. A similar result holds for the BM-WM comparison.

We now determine a prior on  $\beta_3$ . Since  $HR_I = HR_1 HR_2 e^{\beta_3}$ , our prior specification for  $HR_I$  naturally depends on what we thought about  $HR_1$  and  $HR_2$ . Our best estimate of  $\beta_3$  is taken to be  $\beta_{03} = \log[HR_{0I}/(HR_{01}HR_{02})]$ . Moreover, using the independence of the  $\beta_j$ s,

$$\begin{aligned} 0.95 &= \Pr(HR_I \leq u_I | \beta_{01}, \beta_{02}) \\ &= \Pr\left[e^{\beta_3} \leq \frac{u_I}{HR_{01}HR_{02}} | \beta_{01}, \beta_{02}\right] \\ &= \Pr\left[\frac{\beta_3 - \beta_{03}}{\sigma_{03}} \leq \frac{\log[u_I/(HR_{01}HR_{02})] - \log[HR_{0I}/(HR_{01}HR_{02})]}{\sigma_{03}}\right] \\ &= \Phi\left[\log(u_I/HR_{0I})/\sigma_{03}\right]. \end{aligned}$$

We get  $\sigma_{0I} = \log(u_I/HR_{0I})/1.645$ .

Alternatively, we could try to elicit direct information about  $\beta_3$  by eliciting information on the relative sizes of, say, the WF-WM hazard ratio  $e^{\beta_1}$  and the BF-BM hazard ratio  $e^{\beta_1 + \beta_3}$ . The relative size is  $e^{\beta_3} = e^{\beta_1 + \beta_3}/e^{\beta_1}$ . A best guess that the BF-BM hazard ratio is twice as large as the WF-WM hazard ratio corresponds to  $e^{\beta_{03}} = 2$ . One would also need an upper bound on the relative size. We note that our assumption of independence of the  $\beta_j$ s is equivalent to the assumption that  $HR_1$ ,  $HR_2$ , and the ratio of WF-WM and BF-BM hazard ratios are mutually independent.

We can obtain a partial prior by placing informative priors on only a few of the  $\beta_j$ s. The remaining  $\beta_j$ s get diffuse normal prior distributions. If there is an interaction in the model, we could elicit information about fewer relative hazards. For example, in the previous illustration we could place informative priors on  $\beta_1$  and  $\beta_2$ , and use a diffuse  $N(0, \sigma_{03}^2)$  prior on  $\beta_3$ .

**EXERCISE 13.12.** Suppose you have data on time to death from diagnosis of cancer with two dichotomous predictors Sex and Race and you fit the PH model. Let Sex be the first variable, taking the value 1 for a Female and 0 for a Male. Let Race take the value 1 if the person is Black and 0 if

White. (a) With no interaction, construct an informative prior on the two regression coefficients that reflects beliefs that the *HR* comparing Females to Males is 1 with a 5% lower bound of 1/10, and that the *HR* for comparing Blacks to Whites is 3/7 with a 95% upper bound of 3/2. Write WinBUGS code to induce the prior on  $(\beta_1, \beta_2)$ . (b) Now assuming an interaction, construct an informative prior for  $(\beta_1, \beta_3)$  if you have elicited the following information. Assume a best guess for the *HR* comparing a WF to a WM of 5/9 with 95% upper bound 3/2, and a best guess for the *HR* comparing a BF to a BM of 5 with a lower 95% bound of 4/5. Write WinBUGS code to obtain the induced prior on  $(\beta_1, \beta_3)$  (c) Now assume the same elicitation in (b) and add a best guess for the *HR* comparing a BF to a WF of 1.0 with an upper 95% bound of 3. Find the induced prior on  $(\beta_1, \beta_2, \beta_3)$ . Write WinBUGS code to induce the full prior. In the same code, induce the prior on the four *HRs* comparing (i) WF to WM, (ii) BF to BM, (iii) BF to WF, and (iv) BM to WM. (d) Modify the prior in (c) to be only partially informative, using only the first two specifications. Obtain the induced prior on all four *HRs*.

### 13.2.5 Priors for $\lambda$

The baseline hazard is the hazard when  $x = 0$ . It is particularly useful to standardize all continuous variables so that there is a convenient interpretation for the predictor vector  $x = 0$  and thus for the baseline hazard.

In Chapter 12, we discussed a bathtub-shaped hazard that reflects mortality from birth to death. However, selecting actual numbers for the hazard function is difficult without a lot of experience. We want a flexible baseline hazard function that is not overly constrained to any particular shape. However, in its extreme form, too much flexibility can lead to over-fitting and poor prediction. In Chapter 15 we discuss “centering” nonparametric families of distributions on parametric families. This is one way to scale back on over-fitting. The development we consider here “centers” the prior on an Exponential regression model. Centering on other families like the log-normal or Weibull is also possible.

The idea is to place independent Gamma( $e_k, f_k$ ) distributions on the  $\lambda_k$ s. The use of reference priors in which all the  $e_k$ s and  $f_k$ s are small can easily lead to over-fitting. On the other hand, even when an expert is available, it may be difficult to elicit reasonable best guesses and percentiles for every  $\lambda_k$ . Instead, we develop a family of Gamma priors that centers our prior for the baseline hazard on the Exponential distribution.

To center the prior for the baseline hazard on an  $\text{Exp}(\lambda_*)$  distribution, which has constant hazard  $\lambda_*$  for all  $t > 0$ , we take

$$\lambda_k \stackrel{\text{ind}}{\sim} \text{Gamma}(\lambda_* w_k, w_k).$$

Then, for all  $k$  and any  $t$ , we have  $\lambda_* = E(\lambda_k) = E[h_0(t)]$ . We take the development a step further by requiring the prior variance of  $\lambda_k$  to be inversely proportional to the corresponding interval length  $a_{k+1} - a_k$ . This results in  $\lambda_k \sim \text{Gamma}([a_{k+1} - a_k]\lambda_* w, [a_{k+1} - a_k]w)$ . Obviously, we cannot impose this requirement for the infinite interval associated with  $\lambda_K$ , so we simply pick  $\lambda_K \sim \text{Gamma}(\lambda_* w_K, w_K)$ . As discussed earlier, if there are no observations in the last interval, as will be the case in our data model, the prior and posterior on  $\lambda_K$  will be identical and of little interest. For  $0 \leq t < a_K$ , this prior model results in a random baseline cumulative hazard  $H_0(t)$  that is a special case of something called a *Gamma process prior*, cf. Kalbfleisch (1978). The prior involves two tuning (hyper)parameters, namely  $\lambda_*$  and  $w$  (as well as  $w_K$ ). We fix both at constants in our illustrations but prior distributions could be placed on them.

If we let  $w$  be large, then the Gamma distributions will concentrate on  $\lambda_*$  and the baseline hazard will mimic that of an  $\text{Exp}(\lambda_*)$  distribution. Letting  $w$  be small gives a larger prior variance on  $\lambda_k$ , allowing the data to drive departures from the Exponential. The value  $w = 0.001$  might provide a reference value that allows for substantial uncertainty. These comments also apply to  $w_K$ .

A value of  $\lambda_* = 100$  corresponds to a prior expectation that events happen at a high rate across all intervals, while a value of 1/100 indicates that events are somewhat rare, depending on the time

scale. We now provide an ad hoc justification for thinking of  $0.69/\lambda_*$  as the median time of survival under baseline conditions of  $x = 0$ . The random median time of survival  $\tilde{t}$  satisfies  $0.5 = S_0(\tilde{t}) = \exp[-H_0(\tilde{t})]$ , so  $\log(0.5) = -H_0(\tilde{t})$ . Taking the prior expectation of the equation  $\log(0.5) = -H_0(t)$  leads to  $\log(0.5) = -E[H_0(t)] = -\lambda_* t$ . Solving gives  $0.69/\lambda_*$  as a rough approximation to  $E(\tilde{t})$ . Alternatively, we could imagine that the data under baseline conditions of  $x = 0$  were really  $\text{Exp}(\lambda_*)$  so that the median time to event is  $0.69/\lambda_*$ . With this interpretation, if we pick  $\lambda_*$  to be  $1/100$ , the median time to event is 69 units of time. If time is measured in months, that is nearly 6 years. If  $\lambda_*$  is  $1/10$ , the median is about 6.9 months. By eliciting a best guess and a percentile for the median event time under baseline conditions, we could place an informative prior on  $\lambda_*$ .

There are advantages to putting a hyperprior on  $\lambda_*$ . One problem with independent Gamma priors for the components of  $\lambda$  is that we typically expect knowledge about adjacent  $\lambda_j$ s to be related. In particular, we expect the baseline  $h_0$  to be reasonably smooth, so the jumps in the step function model (4) should be relatively small. Placing a prior on the hyperparameter  $\lambda_*$  will induce correlation among the  $\lambda_i$ s and also help with the estimation of survival curves over regions with little data, e.g., the extreme right and left tails of the data. A prior on  $\lambda_*$  serves to center the baseline hazard prior distribution on the *family* of  $\text{Exp}(\lambda_*)$  distributions, where the family is indexed by  $\lambda_* > 0$ . When we give a specific value for  $\lambda_*$ , we are centering the prior on a particular exponential distribution. To allow for more correlation among terms that are closer in time, an autoregressive structure for the  $\lambda_i$ s might be reasonable (see the random walk prior for B-splines in Section 15.2). We do not pursue these ideas further. Readers who are interested should consult Ibrahim, Chen, and Sinha (2001). These authors also expand the centering distribution beyond the Exponential.

**EXERCISE 13.13.** *Full Conditional Distributions.* (a) Derive the forms of the full conditional distributions for  $\beta$  and for  $\lambda_1, \dots, \lambda_K$  assuming at least one death in every interval. (b) Explain how to sample each. (c) Suppose the interval  $[a_k, a_{k+1})$  contains no deaths. Show that with  $\lambda_k \sim \text{Gamma}(e_k, f_k)$ , the full conditional has the form  $\lambda_k \sim \text{Gamma}(e_k, \alpha(\lambda_{(k)}, \beta, D))$  for some function  $\alpha(\cdot)$ . If  $e_k$  is small and  $\alpha(\lambda_{(k)}, \beta, D)$  is large, the distribution focuses on small values. Explain why this might cause computational difficulties. Are computational difficulties more likely to occur when  $k$  is large or small? (d) Obtain the posterior mode for the full conditional of  $\lambda_k$ . Use this to give a possible explanation of why the selection of  $w$  could have a large impact on the posterior.

### 13.2.6 Our Data Model

To control problems with noninformative data and intervals that contain no deaths (cf. Exercise 13.13(c)), we let the  $a_k$ s depend on the data. While it normally would be nonsensical to have a distribution for the data that is functionally dependent on those very same data, we view this more like using the data to select a model, and then analyzing the model as if we hadn't done that. (A sin we have committed frequently.)

Consider the observed death times (uncensored data) *augmented by zero* (provided that all death times are positive) and the largest censored value (provided that value is larger than all the actual death times). Write the ordered distinct *augmented* death times as  $\{d_k : k = 1, \dots, k^*\}$  with  $d_k < d_{k+1}$ .

One choice for a model takes  $K = k^*$  and  $a_k = d_k$ . For the Poisson likelihood this would be a disaster because all of the death times occur at the endpoints of the intervals.

A better choice for our purposes uses intervals that are centered on the augmented death times. Again,  $K = k^*$  but now  $a_1 = 0$  and

$$\begin{aligned} a_k &= (d_k + d_{k+1})/2, \quad k = 2, \dots, K-1 \\ a_K &= d_{k^*} + (d_{k^*} - d_{k^*-1})/2. \end{aligned}$$

Since  $d_1 = 0$ ,  $d_1, d_2 \in [a_1, a_2)$  and thereafter  $d_{k+1} \in [a_k, a_{k+1})$ . One death time occurs in each interval, except there are no observations in  $[a_K, \infty)$  and  $[a_{K-1}, a_K)$  contains  $d_{k^*}$ , which is the larger of the largest "death" time and the largest censoring time.

Yet another choice involves an alternative set of intervals that are centered on the augmented death times. Here  $K = k^* + 1$  and

$$\begin{aligned} a_1 &= 0 \\ a_k &= \frac{d_{k-1} + d_k}{2}, \quad k = 2, \dots, K - 1 \\ a_K &= d_{k^*} + \frac{d_{k^*} - d_{k^*-1}}{2}. \end{aligned}$$

Now  $d_k \in [a_k, a_{k+1})$  and there are no observations in  $[a_K, \infty)$ , but now there are no deaths in  $[0, a_2]$ . We actually implement this choice in what follows.

With no observations in  $[a_K, \infty)$ , based on our earlier discussion of noninformative data, we ignore all issues involving  $\lambda_K$  because  $\lambda_K$  is independent of the other parameters and its prior and posterior are identical.

It is well known in survival analysis that inferences about survival prospects in the right tail of the survival curve are highly uncertain since there are little or no data there. *We recommend not making inferences about survival prospects that are past the largest death time in the data.* Moreover, the modeling assumption that the hazard is constant on the infinite interval  $[a_K, \infty)$  is probably the diciest part of model (4).

Because there are no deaths in  $[0, a_2]$ , we have the computational problems described in Exercise 13.13(c). To overcome this computational difficulty, we essentially fix  $\lambda_1$  at a small number. Note that with no deaths in the interval,  $S_0(a_2) = e^{-\lambda_1 a_2}$  should be near 1 and  $\lambda_1$  near 0. Specifically, we elicit a best guess for  $S_0(a_2)$ , say  $\gamma$  and solve for  $\lambda_1$ , namely  $\lambda_{10} = -\log(\gamma)/a_2$ . Then we pick  $\lambda_1 \sim \text{Gamma}(\lambda_{10} w_1, w_1)$  with  $w_1 = 10,000$  or another very large value. This solves our WinBUGS error message problem for having a shape parameter that is too small at the expense of having to select a value for  $\lambda_{10}$ . One could just select  $\lambda_1$  to be a very small number and not worry about it further. Making inferences about survival times very close to 0 is unusual anyway.

### 13.2.7 WinBUGS Code

The WinBUGS manual provides an illustration of fitting the PH model but it involves substantial differences from our presentation. The first part of our code sets up the expanded data for the Poisson likelihood. It then proceeds to evaluate the likelihood contributions (6). We then include code for constructing survivor functions and hazard ratios plus the prior specifications. With large  $n$  and  $K$ , this code runs considerably slower than previous code we have presented. In particular, the example in Subsection 13.2.9 has  $n = 876$  and  $K = 130$  and it took about five seconds to execute ten iterations of the Gibbs sampler (on the order of eight minutes for every 1,000 iterations).

**EXAMPLE 13.2.1.** *Leukemia Data Revisited.* To illustrate the WinBUGS code, we return to the leukemia study from Example 12.3.1. There are two groups of patients, no censored observations, and time until death from diagnosis with leukemia was recorded for  $n = 33$  patients. Here  $K = 23$  and  $\lambda_{23}$  is ignored. The two groups are  $AG+$  and  $AG-$ . The code below is in regression form using the indicator variable  $AGgroup[i]$ , which takes the value 1 if subject  $i$  is  $AG+$  and 0 if  $AG-$ . With only one predictor, there is only one scalar regression coefficient  $\beta$ . The baseline is  $AG-$  and the hazard for  $AG+$  is  $e^\beta$  times the  $AG-$  hazard. We set  $w = 0.01$ ,  $\lambda_{10} = 0.01$ , and  $\lambda_* = 0.05$ . The latter corresponds to a prior guess that the median time to death is  $20 \times 0.69 = 13.8$  weeks for those who are  $AG-$ .

```
model{
# Set up the a[j]s
# d[1]=0; {d[j]: j = 2,...,k} are distinct death times if
# last observation is a death; else, d[k] is last censored time
# Dimension of vector d is k
a[1] <- 0
```

```

for(i in 2:k){a[i] <- (d[i-1] + d[i])/2}
a[k+1] <- d[k] + (d[k] - d[k-1])/2
# d[j] is in (a[j], a[j+1]), j = 2, ..., k
# Construct H[i,j] and N[i,j]
for(i in 1:n){
  for(j in 1:k){
    HH[i,j] <- step(y[i]-a[j])
    H[i,j] <- HH[i,j]*(min(y[i],a[j+1])-a[j])
    N[i,j] <- HH[i,j]*step(a[j+1]-y[i])*delta[i]
  # Equals 1 iff y[i] is in interval j and delta[i]=1
  }
}
# Likelihood Construction
for(i in 1:n){
  mu[i] <- beta[1]*(AGgroup[i])
  for(j in 1:k){
    N[i,j] ~ dpois(theta[i,j])
    theta[i,j] <- exp(mu[i])*H[i,j]*lam[j]
  }
}
# Prior for lambdas
lamm <- 10000*(0.01) # exp(-a[2]*0.01) = 0.995
lam[1] ~ dgamma(lamm,10000) # lam1 is "fixed" at 0.01
for(j in 2:k){ # prior on baseline hazard bits
  f[j] <- w*(a[j+1]-a[j])
  e[j] <- lamstar*f[j]
  lam[j] ~ dgamma(e[j],f[j])
}
lamstar <- 0.05 # Select value or set prior for lamstar
# lamstar ~ dgamma(a,ww) a=0.05*ww ww=0.01

# Inferences for survival functions
for(j in 1:k){
  dH0[j] <- (a[j+1]-a[j])*lam[j] # (H_0(a[j+1])-H_0(a[j]))
# Survivor function at a[j+1] = exp{-H_0(a[j+1])exp(mu)}
# Surv Fn for AG+ and AG- groups
  S.pos[j] <- pow(exp(-sum(dH0[1:j])), exp(beta[1]))
  S.neg[j] <- pow(exp(-sum(dH0[1:j])), 1)
  S.diff[j] <- S.pos[j] - S.neg[j]
}
# Inferences for the HRs and prior on the betas
for(j in 1:1){
  HR[j] <- exp(beta[j]) # relative hazard
  beta[j] ~ dnorm(0,0.000001)
}
# Data. All data are deaths; k=22 distinct death times
list(n=33, k=23, w=0.01,
      y=c(65,156,100,134,16,108,121,4,39,143,56,26,22,1,1,5,65,
           56,65,17,7,16,22,3,4,2,3,8,4,3,30,4,43),
      AGgroup=c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
                0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0),

```

Table 13.8: *Output from a PH analysis of the leukemia data:  $w = 0.01$ ,  $\lambda_{10} = 0.01$ ,  $\lambda_* = 0.05$ , and  $\Delta S(t) = S_+(t) - S_-(t)$ .*

Node	mean	sd	2.5%	med	97.5%
$HR = e^\beta$	0.29	0.13	0.11	0.27	0.62
$AG-$	$S_-(1.5)$	0.90	0.06	0.75	0.92
	$S_-(12)$	0.44	0.11	0.24	0.44
	$S_-(24)$	0.26	0.10	0.10	0.25
	$S_-(82.5)$	0.04	0.04	<0.001	0.02
$AG+$	$S_+(1.5)$	0.97	0.02	0.92	0.98
	$S_+(12)$	0.79	0.08	0.62	0.80
	$S_+(24)$	0.68	0.10	0.47	0.69
	$S_+(82.5)$	0.34	0.11	0.14	0.34
	$S_+(138.5)$	0.13	0.08	0.02	0.12
$\Delta S$	$\Delta S(1.5)$	0.07	0.05	0.01	0.06
	$\Delta S(12)$	0.35	0.11	0.14	0.35
	$\Delta S(24)$	0.42	0.12	0.17	0.42
	$\Delta S(82.5)$	0.31	0.11	0.11	0.30
	$\Delta S(138.5)$	0.13	0.08	0.02	0.12

```

delta=c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
      1,1,1,1,1,1,1,1,1,1,1,1,1,1,1),
d=c(0,1,2,3,4,5,7,8,16,17,22,26,30,39,43,
    56,65,100,108,121,134,143,156))
# Initial Values. 23 lambdas and 1 beta
list(beta=c(0), # lamstar = 0.05,
  lam=c(0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1,
        0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1))

```

### 13.2.8 Posterior Analysis for Leukemia Data

Table 13.8 gives inferences for the relative hazard  $e^\beta$ ; survivor functions for the  $AG+$  and  $AG-$  groups at 1.5, 12, 24, 82.5, and 138.5 weeks; and their differences at these same times. The times are  $a_3$ ,  $a_9$ ,  $a_{12}$ ,  $a_{18}$ , and  $a_{22}$ . All the 95% posterior intervals for differences in survival exclude 0 indicating survival prospects are statistically better in the  $AG+$  group at each of the times considered. For example, we are 95% sure that the proportion of  $AG+$  patients who will survive at least 24 weeks exceeds the corresponding proportion for  $AG-$  patients by between seventeen and sixty four percentage points. Patients are clearly better off if they are  $AG+$  at diagnosis. The relative hazard comparing  $AG-$  to  $AG+$  is the inverse of  $HR$  and can be estimated as  $1/0.29 = 3.5$ . Similarly, we are 95% sure that the instantaneous failure rate for an  $AG-$  patient is between  $1/0.62 = 1.6$  and  $1/0.11 = 9.1$  times greater than it is for an  $AG+$  patient. Recall from Example 12.3.3 and Table 12.2 that, under the exponential model, the estimated relative median of  $AG+$  to  $AG-$  was 3.5 with a 95% PI of (1.7, 7), *using diffuse priors*. The relative median is the inverse of the relative hazard under an exponential model, so these estimates are also point and interval estimates for the  $HR$  comparing  $AG-$  to  $AG+$  under the exponential model. The exponential results are very similar to those obtained here with the more flexible PH model.

Figure 13.4 plots estimated survival functions from (i) the PH regression, (ii) a two-sample exponential model using independent  $\text{Gamma}(0.0001, 0.0001)$  priors for  $\theta_1$  and  $\theta_2$  (also an AFT model), and (iii) standard frequentist estimates obtained using the `cox.zph` function in R. Since we have not used very informative priors, we would be concerned if our estimates were very different from the well-studied frequentist procedure. From Figure 13.4 we can see that the PH model estimate of

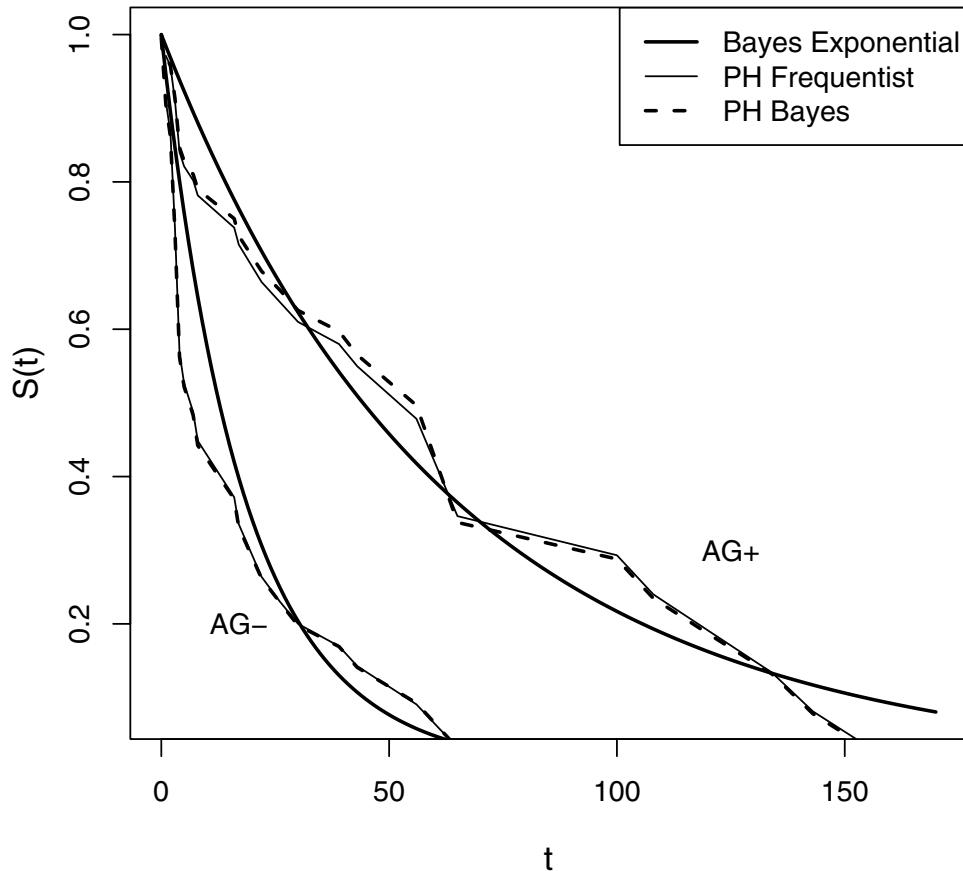


Figure 13.4: *Estimated survival curves for AG+ and AG– leukemia patients.*

the median time to death is about 60 days for AG+ patients and is about 7 days for AG– patients. From Table 12.2, estimated medians from the exponential model are 44 and 13 days for AG+ and AG– patients, respectively.

We performed a sensitivity analysis by changing  $w$  to 0.001, 0.1, and 1. Also we considered  $\lambda_*$  over the range 0.05 to 0.25. We found that, with small values of  $w$ , there was little effect on the final results by changing  $\lambda_*$ . However, with  $w = 1$ , there was an appreciable effect. With  $w = 1$ , increasing  $\lambda_*$  results in survival curve estimates that drop faster than they do with smaller values. Readers should be careful in their choice of  $\lambda_*$  if they intend to use values of  $w$  that are above 0.1. With larger data sets, this effect should be diminished.

### 13.2.9 SAS Analysis of Leukemia Data

Bayesian PH survival models can be fit in SAS, as illustrated here using the leukemia data. In SAS 9.2, a data set that we named `leuk` was created that contains the variables `y`, `delta`, and `AGgroup` exactly as they appear in the WinBUGS code in Section 13.2.7. We then ran the following code to fit a PH model.

```
data groups;
AGgroup=1; output;
AGgroup=0; output;
run;
```

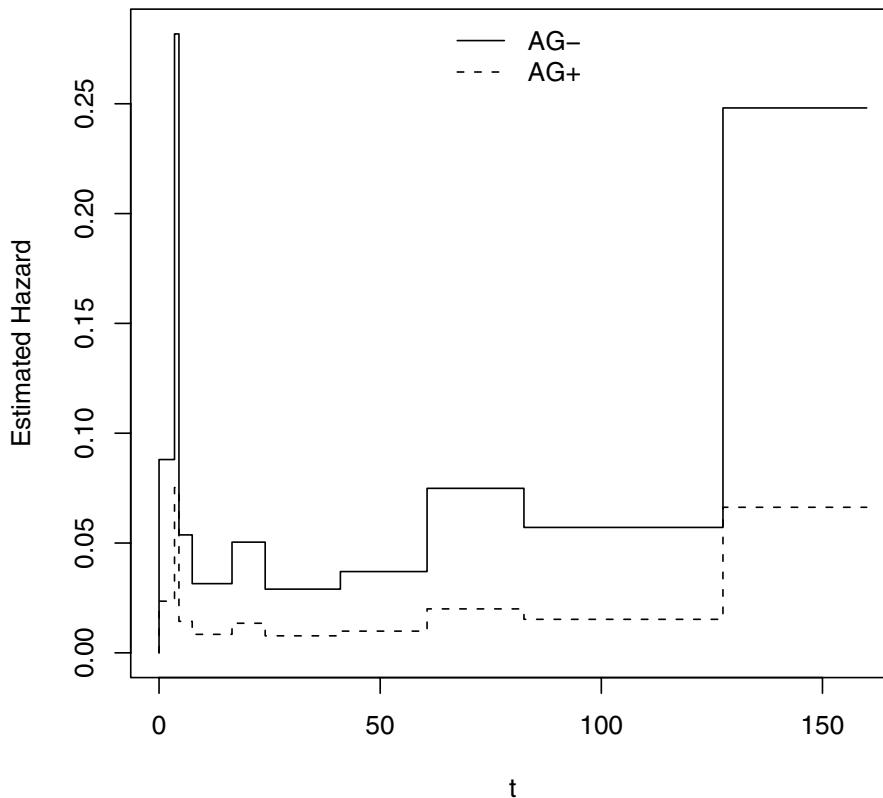


Figure 13.5: Estimated hazard functions from a PH analysis of the leukemia data in SAS.

```

ods pdf; ods graphics on;
proc phreg data=leuk plot(overlay)=survival;
model y*delta(0)=AGgroup;
baseline covariates=groups out=_null_;
bayes nbi=1000 nmc=10000 thinning=5 plots=all
coeffprior=uniform piecewise=hazard (ninterval=10 prior=improper);
quit; ods graphics off; ods pdf close;
* can also do one of several AR(1) models for piecewise hazard;
* e.g. prior=argamma shape=1 scale=10;

```

Unlike in the previous WinBUGS analysis, here we use improper priors, the baseline hazard is approximated by a step function with 10 intervals, and we thin the Gibbs sampler. The posterior mean and standard deviation for the regression coefficient corresponding to the AG+ group are  $-1.32$  and  $0.44$ . From our previous WinBUGS analysis, these values were  $-1.30$  and  $0.43$ . In this model, the baseline hazard  $h_0(t)$  is the hazard for AG- patients, an estimate of which is plotted in Figure 13.5. This figure also contains an approximate Bayesian estimate of the hazard function for AG+ patients where the approximation comes from substituting the posterior mean for  $\beta$  into  $h(t) = e^\beta h_0(t)$ .

### 13.2.10 Another Example

We now use the PH model to analyze data with more observations (including censored observations) and more covariates. The data involve 863 times (in days) to death or censoring after kidney transplantation. The surgeries were performed at the Ohio State University Transplant Center between 1982 and 1992. We use a proportional hazards regression model to investigate differences in survival between the sexes (coded as 1 = Male, 2 = Female) and between races (coded as 1 = White, 2 = Black), while controlling for age in years at the time of transplant. Thus  $x' = (\text{Sex}, \text{Race}, \text{Age})$ , where Age has been standardized by its sample mean and standard deviation. The data are from Klein and Moeschberger (2003, Section 1.7).

There were 140 deaths (16%) over the course of follow-up (about 9.5 years). The number of deaths during the study for the four groups are as follows: among White Males, 73 out of 432 (17%); among Black Males, 14 out of 92 (15%); among White Females, 39 out of 280 (14%); among Black Females, 14 out of 59 (24%). The median (sd) age at transplant was 43.5 (13.2) years for White Males, 47 (13.1) years for Black Males, 41 (14.2) years for White Females, and 44 (12.2) years for Black Females.

We do not present the entire WinBUGS code because some of it (setting up the data and defining the prior on the baseline hazard) is the same code as used for the leukemia analysis. The baseline group was selected as White Males with standardized age 0 (42 years). We set  $w = 0.001$ ,  $\lambda_* = 0.0001$ , and  $\lambda_{10} = 0.01$ . These values were selected without input from a knowledgeable physician. The value  $\lambda_* = 0.0001$  reflects a belief that the median time to death after transplantation under the exponential model would be  $0.69 \times 10000 = 6900$  days, or about 19 years. The following are modifications to the WinBUGS code.

```
# Replace mu[i] with
mu[i] <- beta[1]*(sex[i]-1) + beta[2]*(race[i]-1) +
beta[3]*(age[i]-mean(age[ ]))/sd(age[ ])

# Modify code for survivor functions
# Survivor functions for M and F among whites of average age
S.wm[j] <- pow(exp(-sum(dH0[1:j])), 1)
S.wf[j] <- pow(exp(-sum(dH0[1:j])), exp(beta[1]))
S.wdiff[j] <- S.wm[j] - S.wf[j]
# Survivor functions for M and F among blacks of average age
S.bm[j] <- pow(exp(-sum(dH0[1:j])), exp(beta[2]))
S.bf[j] <- pow(exp(-sum(dH0[1:j])), exp(beta[1]+beta[2]))
S.bdiff[j] <- S.bm[j] - S.bf[j]
# Survivor function for WF aged 2 sd above the mean
S.age.2.wf[j] <- pow(exp(-sum(dH0[1:j])), exp(beta[1]+2*beta[3]))
# Modify code for HRs and prior on beta
for(j in 1:3){
  HR[j] <- exp(beta[j]) # relative hazards
  beta[j] ~ dnorm(0,0.000001)
}
# Modify initial Values
list(beta=c(0,0,0),lam=c(0.1,...,0.1)) # 128 inits for lam
# Data
list(n=863, k=130, w=0.001,
d=c(0,2,3,7,10,17,21,26,28,37,40,43,44,45,50,52,56,
57,59,62,68,69,78,79,88,91,97,98,104,106,119,...,2795,3146))
I[ ] y[ ] delta[ ] sex[ ] race[ ] age[ ]
1     1      0        1       1       46
2     5      0        1       1       51
3     7      1        1       1       55
```

Table 13.9: Output for PH analysis of the kidney transplant data;  $w = 0.001, \lambda_{10} = 0.01, \lambda_* = 0.0001$ .

Node	mean	sd	2.5%	median	97.5%
$e^{\beta_1}(\text{Sex})$	0.93	0.16	0.66	0.91	1.26
$e^{\beta_2}(\text{Race})$	1.05	0.22	0.67	1.03	1.53
$e^{\beta_3}(\text{Age})$	1.9	0.17	1.58	1.89	2.26
$x'$	(F,W,Ave Age + 2sd)				
$S(13.5)$	0.95	0.013	0.92	0.95	0.97
$S(713.5)$	0.70	0.05	0.59	0.70	0.80
$S(2082)$	0.51	0.07	0.36	0.51	0.64
$S(3322.5)$	0.34	0.08	0.18	0.34	0.51
$x'$	(F,W,Ave Age)				
$S(13.5)$	0.986	0.003	0.980	0.987	0.991
$S(713.5)$	0.91	0.015	0.88	0.91	0.93
$S(2082)$	0.83	0.02	0.78	0.83	0.87
$S(3322.5)$	0.74	0.039	0.65	0.74	0.81
$x'$	(M,B,Ave Age)				
$S(13.5)$	0.985	0.004	0.975	0.985	0.991
$S(713.5)$	0.89	0.02	0.85	0.90	0.93
$S(2082)$	0.81	0.04	0.73	0.81	0.87
$S(3322.5)$	0.71	0.05	0.59	0.71	0.81

```

...
862    3211    0        2        2       43
863    3304    0        2        2       52
END

```

The posterior results given in Table 13.9 include the estimated HRs and estimated survival probabilities corresponding to 13.5, 713.5, 2082, and 3322.5 days ( $a_5, a_{74}, a_{115}, a_{130}$ ). There is a clear effect of age that can be seen by comparing estimated survival probabilities for FWs who are two standard deviations above the average age of 42 years at transplant to those who are of average age. Estimated differences are substantial and 95% probability intervals (not shown) for all differences exclude 0. Figure 13.6 presents estimated survival curves comparing these two types of FWs. At the average age of 42 years, estimated survival probabilities don't dip below 70% for any of the 4 groups (MW, FW, MB, FB) over the course of the study (Figure 13.7). Table 13.9 suggests that there is little difference in survival prospects between White Females and Black Males of average age at diagnosis. 95% probability intervals (not shown) for the Sex differences with fixed Race and Age, and for the Race difference for fixed Sex and Age, both include 0.

Consideration of the estimated relative hazards indicates that there is no evidence of a statistically important effect due to Sex or Race since both probability intervals easily contain 1. The intervals are not narrow enough for us to actually conclude that these HRs are essentially 1, but we are confident that they are not too far away from 1. The 95% PIs for the regression coefficients for Sex and Race have 0 squarely in the middle (output not shown). Posterior probabilities that the HRs (regression coefficients) exceed 1 (0) are 0.30, 0.56, and 1.0, respectively, indicating that age at transplantation is an important predictor but that Sex and Race are not. Clearly, being older when the transplant takes place reduces prospects of survival relative to younger transplant patients. We did a sensitivity analysis by setting  $w = 0.01$  and  $\lambda_* = 0.01$  and the results were virtually identical.

With all three predictors, the DIC statistic was 2529. With Sex and Race removed it was 2522.7, indicating that the simpler model is an improvement. With Age as the only predictor in the model, inferences for survival functions at average age were practically identical to those for FW of average

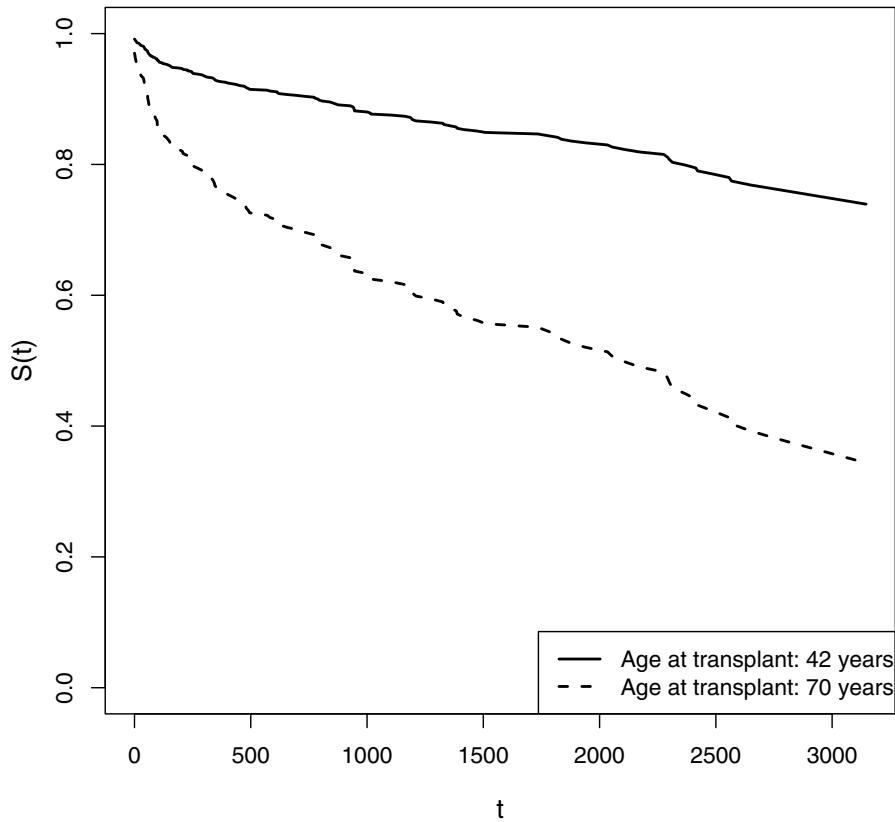


Figure 13.6: Estimated survival curves for white females of average age at transplant and those at 2 standard deviations above average age;  $w = 0.001$ ,  $\lambda^* = 0.0001$ , and  $\lambda_{10} = 0.01$ .

age in Table 13.9; inferences for the *HR* corresponding to Age were also practically identical so we don't reproduce them here.

A (psychotically) careful reader will note that our code only included “death” times in our  $d$  vector, instead of also including the largest censoring time, which was 3434. The data being ignored will make little difference in final results.

Earlier in this section, we alluded to situations where survival curves might cross. Suppose you have two treatments and that survival is better under the first treatment early in the study but better under the second treatment later on. This constitutes an interaction between treatment and time. Define a variable  $\text{Trt}$  with  $\text{Trt} = 1$  for the first treatment and  $\text{Trt} = 0$  for the second. An appropriate model allows for an effect of  $\text{Trt}$  up to a particular point in time, say  $t_c$ , and for a different effect after that. Let  $z(t)$  take the value 0 for  $t < t_c$ , and the value 1 for  $t \geq t_c$ . Then write the model

$$h(t | \text{Trt}, z(t)) = \exp[\beta_1 \text{Trt} + \beta_2 \text{Trt} \times z(t)] h_0(t).$$

Under this model, the *HR* comparing an individual with  $\text{Trt} = 1$  to one with  $\text{Trt} = 0$  is  $e^{\beta_1}$  for  $t < t_c$ , and is  $e^{\beta_1 + \beta_2}$  for  $t \geq t_c$ .

The variable  $z(t)$  is called a (fixed) *time dependent covariate (TDC)*. There also exist random TDCs. For example, the condition of an individual may change during the study, for example eating

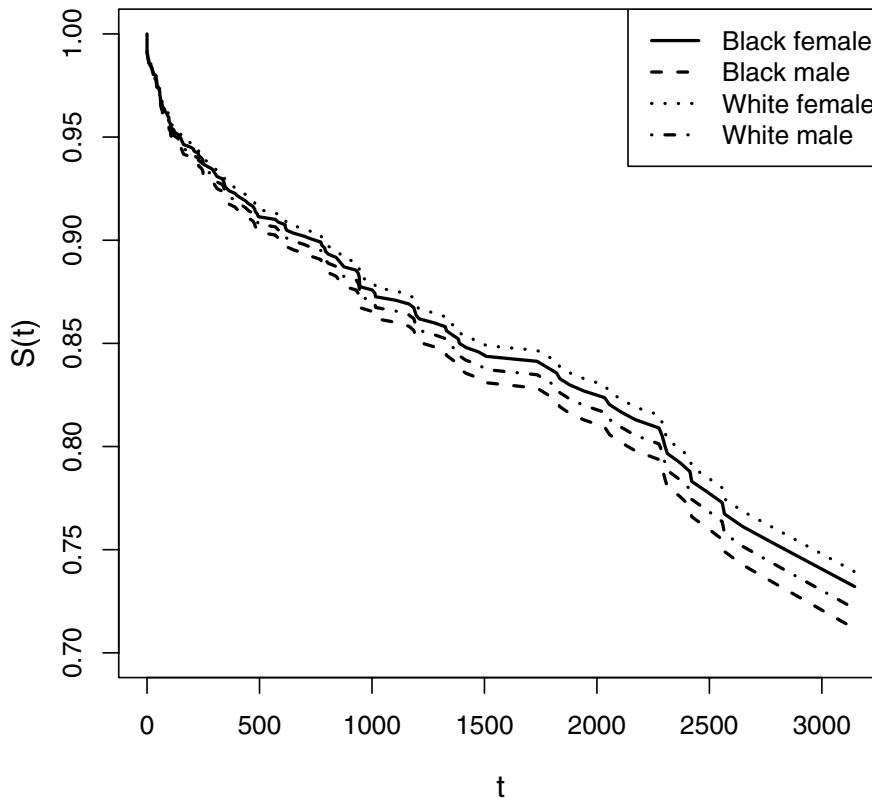


Figure 13.7: Estimated survival curves from a PH analysis of the kidney data;  $w = 0.001$ ,  $\lambda_* = 0.0001$ , and  $\lambda_{10} = 0.01$ .

or smoking habits. Further consideration of time dependent covariates occurs briefly in Section 15.3. Interested readers can consult Collett (2003) or Klein and Moeschberger (2003) for results in frequentist survival analysis, or Hanson, Johnson, and Laud (2009) for semi-parametric Bayesian analyses.

**EXERCISE 13.14. Cow Abortion Data.** Consider the two sample version of the cow abortion data from Section 1.4. The given data are times to natural abortion from conception for two groups; cows that are infected and those that are not. (a) Analyze the data using the PH model. Make inferences for the HR and for survivor probabilities in the two groups. Obtain DIC. (b) Compare with AFT model analyses. (c) Modify the code to place a Gamma( $\tilde{\lambda} v, v$ ) prior on  $\lambda_*$  with  $\tilde{\lambda}$  equal to the value that you used as  $\lambda_*$  in the previous analysis. Use several different weights  $v$  on this prior and compare with your analysis in part (a). Obtain DIC and compare with the results obtained in part (a).

**EXERCISE 13.15. Larynx Cancer Data.** Analyze the larynx cancer data from Examples 13.1.1–13.1.3. Compare a PH analysis with AFT model analyses based on (i) normal, (ii) extreme value, and (iii) logistic errors.

**EXERCISE 13.16. Ovarian Cancer Data.** Analyze the ovarian cancer data available at the book website. The data were taken from Collett (2003, p. 189). They involve the treatment of ovarian

cancer with two different forms of chemotherapy. The data give time to death in days from diagnosis; Treatment 1 = single (chemotherapy) and Treatment 2 = combined (with radiation). See Collett for more details. Compare a PH analysis with an AFT analysis based on (i) normal and (ii) extreme value errors.

**EXERCISE 13.17.** *Kidney Data.* Re-analyze the kidney data to check if there are any interactions among the three predictors: Age, Race, and Sex. Look at one interaction at a time and don't run any code for more than 2,000 iterations since it takes a while to fit the models. Obtain the DIC statistics in each case. Be sure to look at history plots. If any of the interactions seems worth pursuing, give interpretations of what they mean using appropriate relative hazards and survival probabilities.

**EXERCISE 13.18.** *Lung Cancer Data.* (a) Using PH regression, carry out a full analysis of the lung cancer data that were analyzed with a log-normal model in Section 1.6. (b) Analyze the data using log-normal, Weibull, and log-logistic regression models. Select a model from among these, and compare that analysis with your PH analysis in part (a). Code for the log-normal model is given below. When we used as starting values `list(beta=c(0,0,0), tau=1)` we got an error message. Then we used `list(beta=c(10,0,0), tau=1)` and the program ran. We also noticed that there was very high autocorrelation in iterates, so we then standardized age and retried these two starting values with exactly the same results, except that the chains were much better behaved. In any event, we had to try different starting values to get it to run. This is another example of a situation where real prior information would certainly have helped, and where standardizing a continuous covariate can help the behavior of the Markov chain. We leave the code below with unstandardized age so that results will conform with those in Section 1.6.

```
model{
  for(i in 1:121){
    t[i] ~ dlnorm(mu[i], tau)I(c[i],)
    mu[i] <- beta[1] + beta[2]*g[i] + beta[3]*a[i]
  }
  tau ~ dgamma(0.1,0.1)
  for(i in 1:3){beta[i] ~ dnorm(0,0.001)}
}
list(beta=c(8,0,0), tau=1)

g[ ] a[ ] t[ ] c[ ]
0      56    730    0
0      70    NA     1980
0      56    260    0
0      54    NA     1883
0      74    1194   0
0      65    NA     1624
...
1      79    440    0
1      71    251    0
1      63    254    0
END
```

**EXERCISE 13.19.** Modify our WinBUGS code for PH models to handle a single fixed time dependent covariate.

### 13.3 Survival with Random Effects

Consider survival models with random effects. The need for random effects arises just as it did in Chapters 8 and 10 to deal with grouped data. If our sample consists of randomly selected groups (perhaps hospitals) and subsequently selecting individuals from these groups to observe until an event, then we expect observations on individuals within groups to be correlated due to their common environment. For example, we might investigate various treatments' effect on time to death after diagnosis with some terrible disease. Large prospective studies often involve sampling individuals from various medical centers across the country. The data may be in the form we have discussed in Chapters 12 and 13, only now with additional information about group membership. Correlated time to event data is also obtained if we observe recurring events on the same individual through time, e.g., times to recurrent infections in AIDS patients). The data in Section 1.7 on successive times to an armadillo kill for a given hunter are correlated. One way to model correlation is to include a random effect for each individual that is shared through repeated observations on the same individual.

The time to the  $j$ th event for the  $i$ th individual, or the time to event for the  $j$ th individual in the  $i$ th group, say  $T_{ij}$ , can be modeled as

$$\log(T_{ij}) = x'_{ij}\beta + b_i + \sigma\epsilon_{ij}; \quad b_i | \sigma_b \sim N(0, \sigma_b^2) \quad \perp \quad \epsilon_{ij} \sim F_\epsilon \quad (7)$$

for a random effects AFT model, just as in model (10.2.4). Alternatively,

$$h(t | x_{ij}, b_i) = h_0(t)e^{x'_{ij}\beta + b_i}; \quad b_i | \sigma_b \sim N(0, \sigma_b^2) \quad (8)$$

is a Cox PH model with random effects.

There is nothing magical about selecting a normal random effects distribution. We could just as well take  $\gamma_i \equiv e^{b_i} \sim \text{Gamma}(\tau_b, \tau_b)$ , which has mean 1 and precision  $\tau_b$ . (In the survival literature the  $\gamma_i$ s have been called *frailties*, but since the term "random effects" has been around much longer, we stick with that terminology.) We could alternatively place distributions on  $b_i$  or  $\gamma_i$  that have much more flexibility. Natural choices are fixed dimension mixtures of normals or Gammas or versions of the other nonparametric mixtures discussed in Section 15.1. The nice thing about mixtures is that they allow subgroups (or individuals) to share a common random effects distribution that is distinct from those of other subgroups. For instance, in sampling hospitals there may be a particular subgroup of hospitals with better survival prospects than other subgroups. Mixture models may discover that.

As in Chapter 10, we can place a reference prior on the parameters of the AFT models. This would be  $p(\beta, \sigma_b, \sigma) \propto p(\sigma_b)/\sigma$  and be approximated using normal priors with mean 0 and large variances on the regression coefficients, and a Gamma prior with small parameters on  $\sigma$ . We only use proper priors on  $\sigma_b$  so as to avoid the possibility of an improper posterior. For a reference prior, we use a uniform distribution on  $\sigma_b$  over a large range, while we generally use a Gamma distribution when an informative prior on  $\sigma_b$  is constructed. Such priors can also be used for  $\beta$  and  $\sigma_b$  in the PH model with random effects.

Placing informative priors on the regression coefficients for an AFT or PH model with random effects is virtually identical to the procedure without random effects. As in Chapters 8 and 10, elicit information about the median event time or hazard ratios conditional on  $b_i = 0$  (i.e., for a typical group or individual). Priors on  $\sigma$  and  $\sigma_b$  for the AFT model are determined just as in Section 1.7, only now conditional on a particular covariate combination  $x_*$ . For  $\sigma$ , we think about a percentile of the failure distribution for "typical" individuals with this covariate combination and with  $b = 0$ . Then, the  $1 - \alpha$  percentile of failure times for such an individual is  $e^{x'_*\beta + \sigma w_\alpha}$ . We have already specified a prior guess for  $\beta$ , say  $\beta_0$ , and our elicitation for  $\sigma$  is conditional on  $\beta = \beta_0$  with  $\beta \perp \sigma$ . Letting  $M = e^{x'_*\beta_0}$  as in Section 1.7, we can proceed to find a best guess and a percentile for  $\sigma$ , exactly as was done there. The prior on  $\sigma_b$  for the AFT model is obtained exactly as in Section 1.7. Priors for  $\sigma_b$  in the PH model are just slightly different than for the AFT model. For the AFT

model, we were thinking about an acceleration factor  $e^\delta$ . This could also be regarded as a relative median, comparing two individuals with the same covariate combination, but the denominator has the median for a “typical” individual and the numerator has the median for a randomly selected individual. For the PH model, consider the hazard ratio comparing the same individuals.

**EXERCISE 13.20.** The WinBUGS code below is for a log-normal AFT model for the Ache “search time” hunting data with random effects and no covariates. The data (available from our website) include the two covariates mentioned in Section 1.7. (a) Run a program to verify the results presented in Section 1.7. (b) Modify the code so that you can estimate the survival function for a “typical” hunter, and for hunters who are at the 90th percentile of hunters. Plot the two on the same graph. (c) Modify the code so that you can handle the two covariates. You will need a prior for the three regression coefficients. Using all of the information from Section 1.7, construct an appropriate partial prior for the regression coefficients (assuming you don’t have prior information about how the two covariates will affect the response). Analyze the data and decide whether or not to keep the covariates in the model based on your exploration.

```

model{
  for(i in 1:324){
    SEARCH[i] ~ dlnorm(mean[i],tau)I(LOWER[i],)
    mean[i] <- mu + delta[ID[i]]
  }
  for(i in 1:15){delta[i]~dnorm(0,tau.d)} #delta[15] is predictive
  mu ~ dnorm(4.5, 10)
  tau <- 1/(sigma*sigma)
  tau.d <- 1/(sigma.d*sigma.d)
  sigma ~ dgamma(2.29, 2.92)
  sigma.d ~ dgamma(8.1, 24.5)
  mean.pred <- mu + delta[15]
  t.pred ~ dlnorm(mean.pred,tau)
  # covariates in the data list not used in simple analysis
  junk <- FIRSTLAST[1]+AGE[1]
}
list(mu=4, sigma=1, sigma.d=1)
ID[ ]   SEARCH[ ]   LOWER[ ]   FIRSTLAST[ ]   AGE[ ]
1       NA          13         0            63
1       NA          2          0            63
1       9           0          0            63
1       37          0          0            63
...
14      NA          1          0            56
14      NA          6          0            56
14      11          0          0            56
14      9           0          0            56
END

```

---

## Chapter 14

---

# Binary Diagnostic Tests

---

In this chapter, we illustrate Bayesian models and methods for two commonly encountered tasks in medicine: (i) evaluating the performance of diagnostic tests and (ii) the use of diagnostic test data to estimate disease prevalence. We define a *diagnostic test* as any instrument that provides information about the presence or absence of some condition of interest, referred to as the “disease.” Testing accuracy is assessed by estimating the sensitivity and specificity. In this chapter, all methods are based on test data with dichotomous outcomes. The material presented is taken largely from work that is catalogued at the website [www.epi.ucdavis.edu/diagnostictests/](http://www.epi.ucdavis.edu/diagnostictests/), which contains papers and WinBUGS code that apply to this chapter and beyond.

Diagnostic tests discriminate (imperfectly) between diseased and non-diseased individuals. Such tests include diagnostic imaging, virus isolation, serology, bacterial culture, and microarray technology for identifying genes that are correlated with disease. Diagnostic tests can provide information on binary (+/−), ordinal (ordered categorical responses), or continuous scales. Binary tests include presence or absence of clinical signs, genetic predisposition to disease, or whether a virus was isolated. Radiologists often use a 5-point ordinal scale to classify cancer status from X-ray results and ordinal data also come from titrations using increasingly diluted volumes of biological samples. Two common examples of continuous tests are procedures for measuring antibodies to infectious agents and measuring biomarker concentration levels such as CD-4 in HIV patients.

Diagnostic tests are generally imperfect. Two types of misclassification are possible in binary tests. Diseased individuals can test negative and non-diseased individuals can test positive. The implications of false negative errors can be life-threatening, with diseased individuals failing to obtain prompt treatment. A false positive test may result in the physical, emotional, and financial burdens of further testing or even unnecessary treatment. Accuracy of diagnostic procedures must be established before tests can be marketed for widespread use. We focus on a test’s ability to correctly predict (i.e., diagnose) the disease status of individuals. Other economic and logistical factors contribute to the complete evaluation of a medical test.

The texts by Pepe (2003) and Broemeling (2007) address frequentist and Bayesian approaches, respectively, to test evaluation. Both of these authors focus primarily on methods in which a perfect, definitive, but often expensive or inconvenient *gold-standard* test is available to ascertain true disease status. While we consider methods based on known disease status, we primarily focus on test evaluation when we don’t know who is or is not diseased. Further information on the ideas presented here can be found in Hui and Walter (1980), Joseph, Gyorkos, and Coupal (1995), Johnson, Gastwirth, and Pearson (2001), Johnson, Su, Gardner, and Christensen (2004), and Branscum, Gardner, and Johnson (2004, 2005), as well as at the UC Davis website [www.epi.ucdavis.edu/diagnostictests/](http://www.epi.ucdavis.edu/diagnostictests/).

Section 1 introduces basic ideas. Section 2 examines situations with one test and one population. Section 3 considers two tests and two populations. The chapter concludes with a discussion of *prevalence distributions*.

### 14.1 Basic Ideas

Binary diagnostic tests result in either a positive test outcome, denoted  $T^+$ , or a negative test outcome, denoted  $T^-$ .  $D$  denotes presence of the disease and  $\bar{D}$  indicates its absence. Three parameters are typically used to characterize a diagnostic test. Two of these, sensitivity and specificity, are measures of test accuracy. The third parameter is the prevalence of the disease. Two other measures, predictive values positive and negative, address the clinical utility of the test. A simple test was discussed in Example 2.1.1.

**EXAMPLE 14.1.1.** *Coronary Artery Disease.* Weiner et al. (1979) and Pepe (2003) examined data from a study that used results on an exercise stress test (EST  $\equiv T$ ) to diagnose coronary artery disease (CAD  $\equiv D$ ). The data have a multinomial distribution with  $n = 1465$ .

Data	CAD	$\bar{CAD}$	Total
EST $^+$	815	115	930
EST $^-$	208	327	535
Total	1023	442	1465

To observe such data requires the existence of a gold standard that can unambiguously identify individuals as diseased or non-diseased.

We begin by defining terminology along with some elementary non-Bayesian calculations from the example. The *sensitivity* of a test, denoted  $\eta$ , is the probability that a diseased individual tests positive, thus

$$\eta \equiv \Pr(T^+|D).$$

From the example, an obvious estimate of the sensitivity is  $\hat{\eta} = 815/1023 \doteq 0.8$ . The *specificity*, denoted  $\theta$ , is the probability that a non-diseased individual tests negative, so

$$\theta \equiv \Pr(T^-|\bar{D}).$$

From the example, an estimate of the specificity is  $\hat{\theta} = 327/442 \doteq 0.8$ . From a practical point of view, it is nonsense to have  $\theta < 0.5$ . If  $\theta < 0.5$  we should reverse the meaning of what constitutes a negative test result. The disease *prevalence* in the source population is

$$\pi \equiv \Pr(D).$$

The data in Example 14.1.1 provide a random sample of “at risk” men, so we estimate prevalence as  $\hat{\pi} = 1023/1465 \doteq 0.7$ , much higher than the general population. In practice, one might only consider the information about test accuracy to be of interest to the general population.

A test’s clinical utility is measured by its predictive accuracy, which depends on sensitivity, specificity, and prevalence. The conditional probability of  $D$  for those who test positive, called the *predictive value positive* (PVP), is obtained using Bayes’ Theorem:

$$\text{PVP} = \Pr(D|T^+) = \frac{\Pr(T^+|D)\Pr(D)}{\Pr(T^+)} = \frac{\eta\pi}{\eta\pi + (1-\theta)(1-\pi)}. \quad (1)$$

The law of total probability is used to obtain

$$\Pr(T^+) = \Pr(T^+|D)\Pr(D) + \Pr(T^+|\bar{D})\Pr(\bar{D}) = \eta\pi + (1-\theta)(1-\pi).$$

We could estimate PVP using Bayes’ Theorem and our previous estimates

$$\widehat{\text{PVP}} = \frac{.8(.7)}{.8(.7) + (1-.8)(1-.7)} = .88,$$

where the reported estimates of  $\eta$ ,  $\theta$ , and  $\pi$  were all rounded off. Equivalently, we can estimate PVP directly as  $815/930 = 0.88$ . The *predictive value negative (PVN)* is the probability of being disease negative for those who test negative:

$$\text{PVN} = \Pr(\bar{D}|T^-) = \frac{\Pr(T^-|\bar{D})\Pr(\bar{D})}{\Pr(T^-)} = \frac{\theta(1-\pi)}{\theta(1-\pi)+(1-\eta)\pi}. \quad (2)$$

An estimate of PVN from the example is  $327/535 = 0.6$ . These measures also go by the names *positive predictive value* and *negative predictive value*.

**EXERCISE 14.1.** Prove that the conditional distribution of the number of true positives out of the number of  $EST^+$  individuals is  $\text{Bin}(930, \text{PVP})$ , and, independently, the number of true negatives out of the number of  $EST^-$  individuals is  $\text{Bin}(535, \text{PVN})$ .

Two other measures of test performance are the positive and negative diagnostic likelihood ratios, which are defined as

$$\text{DLR}^+ = \frac{\eta}{1-\theta} \quad \text{and} \quad \text{DLR}^- = \frac{1-\eta}{\theta}.$$

$\text{DLR}^+$  is the ratio of the probability of a true positive to the probability of a false positive. This increases towards  $+\infty$  as classification accuracy increases. An estimate from the example is  $[815/1023]/[115/442] = 3$ . A positive test result is about 3 times more likely to come from a diseased person. On the other hand,  $\text{DLR}^-$ , the ratio of false negative to true negative probabilities, equals 0 for a perfect test. From the data, an estimate is  $[208/1023]/[327/442] = 0.27$ , so a negative test result is about one fourth as likely to come from a diseased person. Finally,  $\text{DLR}^+/\text{DLR}^-$  gives the odds of a positive test result from a diseased person relative to the odds of a positive test result from a non-diseased person, i.e., the odds ratio.

With multinomial sampling, every entry in the data table and every marginal total has a binomial distribution, so all of the naive estimates are maximum likelihood estimates. For example, the number of individuals who have CAD is  $\text{Bin}(n, \pi)$  with  $n = 1465$ . The number of true positives, namely the number out of 1465 who are both  $EST^+$  and have CAD is  $\text{Bin}(n, \eta\pi)$ .

The sampling scheme for the CAD data is *cross-sectional*. It is one sample from the population of interest collected at a particular point in time: a cross-section of the population. With this sampling scheme, we are able to estimate all parameters.

From a cross-sectional sample, studying a rare disease like ALS (Lou Gehrig's disease) is difficult. The sample size would have to be huge to get a reasonable number of diseased patients into the study. *Case-control* sampling involves taking independent binomial samples of both diseased and non-diseased patients. This easily provides reasonable numbers for both groups but does so at the cost of losing information on the prevalence of the disease in the population being sampled. This sampling scheme is sometimes called *product binomial* where "product" indicates that the samples are independent. Both the number of cases (diseased individuals) and the number of controls (non-diseased individuals) are determined by the sampling scheme. Case-control studies were discussed in Subsection 5.1.3.

Both cross-sectional and case-control sampling presuppose the existence of a gold standard, some absolute method of determining disease status (one that is not subject to the vicissitudes of the diagnostic test in question). They also require identifying appropriate populations for sampling. A common sampling scheme is neither cross-sectional nor case-control but involves one sample from the diseased population and another sample from the population of interest that contains a mixture of both diseased and non-diseased individuals. This presupposes that you can find sick people who are representative of the entire population of sick people.

It remains to place prior distributions on the parameters. Throughout this chapter we assume that scientific information about sensitivities and specificities can be modeled as

$$\eta \sim \text{Beta}(a_\eta, b_\eta) \quad \perp\!\!\!\perp \quad \theta \sim \text{Beta}(a_\theta, b_\theta).$$

Information about prevalence is also taken to be independent of that for sensitivity and specificity, and we again use a Beta prior

$$\pi \sim \text{Beta}(a_\pi, b_\pi).$$

Subsection 5.1.1 discusses methods for choosing priors for binomial parameters.

**EXERCISE 14.2.** Give independent Beta priors on  $(\pi, \eta, \theta)$  that have modes  $(0.6, 0.9, 0.7)$  and 10th percentiles  $(0.4, 0.6, 0.5)$ , respectively. Use WinBUGS to find the induced priors on  $(PVP, PVN, DLR^+, DLR^-)$ .

## 14.2 One Test, One Population

Consider a single binary diagnostic test that is applied to a random sample of  $n$  individuals from a population. One of two scenarios apply. The first is similar to Example 14.1.1. All individuals are subjected both to the diagnostic test,  $T$ , and to a gold standard (GS) test that determines the disease status with certainty. In the second scenario, only the test  $T$  is applied, so it is known as the no gold standard (NGS) case. We discuss these in separate subsections but we apply a common notation to both.

There are two useful ways to represent the data. First, consider a  $2 \times 2$  table cross-classified by the test result and disease status, e.g., the table in Example 14.1.1. The outcome for the  $i$ th test result and the  $j$ th disease status is denoted  $y_{ij}$  and the entire data are denoted  $Y = \{y_{ij} : i = 1, 2; j = 1, 2\}$ . In the GS case, we observe the entire table, but in the NGS case, we observe only the row totals  $\{y_{i\cdot} : i = 1, 2\}$  where  $y_{i\cdot} = \sum_{j=1}^2 y_{ij}$ . Define the probability corresponding to the  $i$ th row and  $j$ th column as  $p_{ij}$ , which is determined by the sensitivity, specificity, and prevalence.

Probabilities	$D$	$\bar{D}$	
$T^+$	$p_{11} = \eta\pi$	$p_{12} = (1 - \theta)(1 - \pi)$	$p_{1\cdot} = \Pr(T^+)$
$T^-$	$p_{21} = (1 - \eta)\pi$	$p_{22} = \theta(1 - \pi)$	$p_{2\cdot} = \Pr(T^-)$
	$p_{\cdot 1} = \pi$	$p_{\cdot 2} = 1 - \pi$	

Note that

$$\eta \equiv \frac{p_{11}}{p_{\cdot 1}} \quad \text{and} \quad \theta \equiv \frac{p_{22}}{p_{\cdot 2}}.$$

Regardless of whether we get to observe all of  $Y$ ,

$$Y \sim \text{Mult}(n, p) \quad \text{where} \quad p = \{p_{ij} : i = 1, 2; j = 1, 2\}.$$

An alternative notation is convenient when studying the NGS case. In place of  $y_{11}, y_{12}, y_{21}, y_{22}$ , use  $n, y, z_1, z_2$ .

Data	$D$	$\bar{D}$	
$T^+$	$z_1 \equiv y_{11}$	$y - z_1 = y_{12}$	$y \equiv y_{1\cdot}$
$T^-$	$n - y - z_2 = y_{21}$	$z_2 \equiv y_{22}$	$n - y = y_{\cdot 2}$
	$n_D \equiv y_{\cdot 1}$	$n_{\bar{D}} \equiv y_{\cdot 2}$	$n \equiv y_{..}$

In the NGS case, the  $z_j$ s are not observable, only  $y$  and  $n$ . Here  $z_1$  is the number of *true positives* (*TP*),  $y - z_1$  is the number of *false positives* (*FP*),  $z_2$  is the number of *true negatives* (*TN*), and  $n - y - z_2$  is the number of *false negatives* (*FN*). The sample size is  $n$ , the random number of people testing positive is  $y$ , and the numbers of diseased and non-diseased individuals are  $n_D$  and  $n_{\bar{D}}$ , respectively. The four numbers in the body of this table (the  $y_{ij}$ s) are only observable when a GS exists to unambiguously identify diseased and non-diseased individuals.

Because  $n$  is known, the data  $Y$  are determined by  $z_1, z_2$ , and any one of the four marginal totals. Thus  $Y$  is equivalent to the data  $(z_1, z_2, n_D)$ , and also to the data  $(y, z_1, z_2)$ . The *complete*

*data* likelihood is necessary for analyzing the GS case and is helpful in the NGS case. Using the  $Y$  notation, the multinomial likelihood is

$$L(p|Y) \propto \{p_{11}\}^{y_{11}} \{p_{12}\}^{y_{12}} \{p_{21}\}^{y_{21}} \{p_{22}\}^{y_{22}}.$$

Equivalently,

$$L(\pi, \eta, \theta|Y) \propto \{\eta\pi\}^{y_{11}} \{(1-\theta)(1-\pi)\}^{y_{12}} \{(1-\eta)\pi\}^{y_{21}} \{\theta(1-\pi)\}^{y_{22}}.$$

With the alternative data notation, the likelihood becomes

$$L(\pi, \eta, \theta|z_1, z_2, n_D) \propto [\pi^{n_D} (1-\pi)^{n_{\bar{D}}}] [\eta^{z_1} (1-\eta)^{n_D - z_1}] [\theta^{z_2} (1-\theta)^{n_{\bar{D}} - z_2}]. \quad (1)$$

An advantage to the  $z_1$ ,  $z_2$ , and  $n_D$  notation is that each term relates directly to one of the parameters  $\eta$ ,  $\theta$ ,  $\pi$ . Recall that  $n$  is fixed and  $n_D + n_{\bar{D}} = n$ , so knowing  $n_D$  is equivalent to knowing  $n_{\bar{D}}$ . Observe that

$$z_1|\eta, n_D \sim \text{Bin}(n_D, \eta) \quad \perp \!\!\! \perp \quad z_2|\theta, n_D \sim \text{Bin}(n_{\bar{D}}, \theta)$$

and

$$n_D|\pi \sim \text{Bin}(n, \pi).$$

Thus, the joint probability function factorizes as

$$f(n_D, z_1, z_2) = f_0(n_D) f_1(z_1|n_D) f_2(z_2|n_D).$$

Each  $f_j$  is a binomial density. Multiplying them gives the likelihood in (1).

EXERCISE 14.3. Show that (1) is appropriate by writing each  $y_{ij}$  in terms of  $z_1$ ,  $z_2$ , and  $n_D$ .

#### 14.2.1 Gold-Standard Data

With cross-sectional GS data,  $Y$ , or equivalently  $(y, z_1, z_2)$  or  $(z_1, z_2, n_D)$ , is completely observed. Using independent Beta priors on  $\eta$ ,  $\theta$ , and  $\pi$ , inference proceeds separately for each parameter using the methods for binomial data presented in Example 2.3.1, Example 3.1.3, and Section 5.1. This follows from the fact that with the independent Beta priors given at the end of Section 1, the joint posterior density  $p(\eta, \theta, \pi|y, z_1, z_2)$  factors as the product of independent Beta densities:

$$\begin{aligned} \pi|y, z_1, z_2 &\sim \text{Beta}(n_D + a_\pi, n_{\bar{D}} + b_\pi) \\ \eta|y, z_1, z_2 &\sim \text{Beta}(z_1 + a_\eta, n_D - z_1 + b_\eta) = \text{Beta}(z_1 + a_\eta, n - y - z_2 + b_\eta) \\ \theta|y, z_1, z_2 &\sim \text{Beta}(z_2 + a_\theta, n_{\bar{D}} - z_2 + b_\theta) = \text{Beta}(z_2 + a_\theta, y - z_1 + b_\theta). \end{aligned} \quad (2)$$

EXERCISE 14.4. Show that the likelihood derived from  $f(y, z_1, z_2)$  is the same as (1). Then show that the joint posterior based on this likelihood and independent Beta priors results in

$$p(\pi, \eta, \theta|y, z_1, z_2) = p(\pi|y, z_1, z_2) p(\eta|y, z_1, z_2) p(\theta|y, z_1, z_2),$$

where these densities are all Beta with the parameters given earlier.

We have completely characterized the joint posterior and it has a particularly simple form in the GS case. This simplicity does not include computing the PVP and PVN; there is no simple form for their marginal posteriors. However, we can use MCMC methods to sample from the posterior distribution of  $(\eta, \theta, \pi)$  and substitute the simulated iterates into equations (14.1.1) and (14.1.2) to obtain a posterior sample for PVP and PVN.

Table 14.1: Posterior measures of EST accuracy and diagnosability of CAD.

Parameter	Mean	Median	SD	95% PI
$\eta$	0.796	0.796	0.013	(0.771, 0.820)
$\theta$	0.739	0.739	0.021	(0.697, 0.779)
$\pi$	0.698	0.698	0.012	(0.674, 0.721)
PVP	0.878	0.876	0.011	(0.854, 0.896)
PVN	0.611	0.611	0.021	(0.569, 0.651)
DLR <sup>+</sup>	3.067	3.051	0.252	(2.621, 3.608)
DLR <sup>-</sup>	0.276	0.276	0.019	(0.241, 0.314)

EXAMPLE 14.2.1. *CAD*. In the absence of specialist knowledge about the sensitivity, specificity, or prevalence, reference  $U[0,1]$  priors can be placed on  $\eta$ ,  $\theta$ , and  $\pi$ . If a scientist is involved, it would be rare that they did not have some information such as: the prevalence is below 0.5 and the test accuracies are above 0.5. These statements are largely inconsistent with the uniform priors but in the absence of *any* such knowledge, we proceed with uniform priors.

Table 14.1 presents posterior estimates for sensitivity, specificity, prevalence, PVP, PVN, and DLRs. The sensitivity and specificity are moderate, with posterior means and medians of 80% and 74%, respectively. With the uniform priors and the substantial sample sizes, the point estimates are nearly identical to the naive estimates given earlier. Interval estimates are also provided.

The usefulness (diagnosability) of the EST test in this population can be quantified by PVP and PVN. Clinicians can be between 85% and 90% certain that patients who are EST<sup>+</sup> have CAD. The PVN is lower, between 57% and 65%.

EXERCISE 14.5. Use the following WinBUGS code to reproduce Table 14.1.

```
model{
  z1 ~ dbin(eta, nD)
  z2 ~ dbin(theta, nDbar)
  nD ~ dbin(pi, n)
  PVP <- eta*pi / (eta*pi + (1-theta)*(1-pi))
  PVN <- theta*(1-pi) / (theta*(1-pi) + (1-eta)*pi)
  DLRp <- eta / (1-theta)
  DLRn <- (1-eta) / theta
  eta ~ dunif(0,1)
  theta ~ dunif(0,1)
  pi ~ dunif(0,1)
}
list(z1=815, z2=327, nD=1023, nDbar=442, n=1465) # Data
list(eta=0.5, theta=0.5) # Initial Values
```

The use of multinomial or binomial distributions here is really just an approximation. These models assume that data are sampled independently from an infinite population or sampled with replacement. Real populations are almost invariably finite and sampled without replacement. The binomial (multinomial) approximation is good whenever the population size is large and the sample size is relatively small. In some cases, notably animal husbandry, the sample size can be a substantial portion of the population size. In such cases, the binomial distributions are replaced by hypergeometric distributions, see Wilks (1962, Sec. 6.1). We do not go into detail about handling these situations, but Johnson et al. (2004) developed a graphical user interface (GUI) that handles testing problems for sampling either with or without replacement. The GUI can be found at [www.epi.ucdavis.edu/diagnostictests/](http://www.epi.ucdavis.edu/diagnostictests/).

EXERCISE 14.6. Analyze the CAD data by augmenting the WinBUGS code below based on the data  $Y$ , and using uniform priors on the test accuracies and the prevalence. Make inferences about all parameters. Compare your results with those in Table 14.1, which were based on using  $(y, z_1, z_2)$ . They should be the same for all parameters. Pay attention to convergence of the Markov chains.

```
model{
  y[1:2,1:2] ~ dmulti(p[1:2,1:2],n)
  p[1,1] <- eta*pi
  p[1,2] <- (1-theta)*(1-pi)
  p[2,1] <- (1-eta)*pi
  p[2,2] <- theta*(1-pi)
  eta ~ dbeta(a[1],b[1])
  theta ~ dbeta(a[2],b[2])
  pi ~ dbeta(a[3],b[3])
}
y[,1] y[,2]
815 115
208 327
END
list(n=1465, a=c(1,1,1), b=c(1,1,1))
list(pi=0.5, eta=0.5, theta=0.5)
```

EXERCISE 14.7. Write WinBUGS code to analyze the CAD data assuming that the column totals are fixed, that is, assuming case-control data. Use uniform priors as before. Compare inferences for  $\eta$  and  $\theta$  with previous results; they should only differ in Monte Carlo error. Monitor the prevalence  $\pi$  and comment. Also make inferences about the PVP and PVN, but now using a Beta(10,10) prior for the prevalence. There is no information in the data for  $\pi$ . See what happens when you change your prior for the prevalence to a Beta(1,10).

#### 14.2.2 No Gold-Standard Data

In the absence of a gold standard, the problem becomes more challenging. Both  $z_1$  and  $z_2$  are unobservable and instead we observe only  $y$ , the number of individuals out of  $n$  who test positive for the disease. The *population apparent prevalence (PAP)* is given by  $\Pr(T^+) = p_{1\cdot} = \eta\pi + (1 - \theta)(1 - \pi)$  and our observable random variable  $y$  is

$$y|\pi, \eta, \theta \sim \text{Bin}(n, \eta\pi + (1 - \theta)(1 - \pi)).$$

The likelihood function is

$$L(\pi, \eta, \theta|y) \propto \{\eta\pi + (1 - \theta)(1 - \pi)\}^y \{(1 - \eta)\pi + \theta(1 - \pi)\}^{n-y}.$$

The observed datum  $y$  only provides information on the PAP and transformations of it. In other words, the population apparent prevalence is the only identifiable parameter, cf. Example 4.14.2. The sample apparent prevalence,  $y/n$ , is the usual frequentist estimator of it.

Even the PAP has a dubious relation to the other parameters because if we reverse the definitions of everything, that is, if we define

$$\pi_* = 1 - \pi, \quad \eta_* = 1 - \theta, \quad \theta_* = 1 - \eta,$$

we get the same PAP. However, as discussed earlier, it makes no sense to have specificity less than 0.5 and there is little chance that both parameterizations would have an appropriate specificity.

Inferences for  $\eta$ ,  $\theta$ , and  $\pi$  cannot be made unless extra information is available from prior distributions or possibly from other data. As a result, this study design is typically employed only to

Table 14.2: Posterior summaries for  $\pi$ .

y	Beta(1,1) Prior			Beta(1,5) Prior		
	Pr( $\pi \leq 0.05 y$ )	med.	95% PI	Pr( $\pi \leq 0.05 y$ )	med.	95% PI
10	0.006	0.211	(0.080, 0.431)	0.014	0.180	(0.063, 0.353)
5	0.179	0.097	(0.013, 0.256)	0.242	0.083	(0.010, 0.218)
2	0.601	0.040	(0.002, 0.150)	0.656	0.035	(0.002, 0.133)
1	0.760	0.026	(0.001, 0.120)	0.800	0.023	(0.001, 0.106)
0	0.883	0.016	(0.001, 0.089)	0.904	0.015	(0.000, 0.079)

study prevalence using diagnostic tests that have well-established operating characteristics (sensitivity and specificity). The information about  $\eta$  and  $\theta$  typically comes from manufacturers and expert opinion. The analysis is very sensitive to the prior distributions used, so one hopes for a consensus of expert opinion. Be forewarned that manufacturer information is proprietary so in some instances it can be difficult for mere mortals to elicit accurate and precise prior information.

Historically,  $\eta$  and  $\theta$  have been assumed to be fixed and known so that a method of moments estimator for  $\pi$  is obtained by solving  $y/n = \eta\pi + (1 - \theta)(1 - \pi)$  for  $\pi$ . This method runs into trouble if the values for the test accuracies are not consistent with the apparent prevalence  $y/n$ . For example if  $1 - \theta$  and  $\eta$  are specified as 0.2, 0.8, respectively, the apparent prevalence is a weighted average of these numbers, so it must be between them. If the estimated apparent prevalence is, say,  $y/n = 0.9$ , no valid probability  $\pi$  solves the method of moments equation.

The joint posterior distribution of  $\eta$ ,  $\theta$ , and  $\pi$  depends on the data only through the posterior for the apparent prevalence, i.e.,

$$p(\eta, \theta, \pi|y) = p(\eta, \theta, \pi|PAP)p(PAP|y).$$

With independent Beta priors for  $\pi$ ,  $\eta$ , and  $\theta$ , the posterior is

$$p(\pi, \eta, \theta|y) \propto \{\eta\pi + (1 - \theta)(1 - \pi)\}^y \{(1 - \eta)\pi + \theta(1 - \pi)\}^{n-y} \times \pi^{a\pi-1} (1 - \pi)^{b\pi-1} \eta^{a\eta-1} (1 - \eta)^{b\eta-1} \theta^{a\theta-1} (1 - \theta)^{b\theta-1}.$$

Because  $p(\pi, \eta, \theta|y)$  is analytically intractable, we turn to simulation.

**EXAMPLE 14.2.2. Paratuberculosis.** We estimate the prevalence of paratuberculosis in a herd of dairy cows using an imperfect ELISA (enzyme-linked immunosorbent assay) test. Our expert's best guess for  $\eta$  is 0.45 and the expert is 90% sure that  $\eta < 0.55$ . Our expert's best guess for  $\theta$  is 0.99 and is 90% sure that  $\theta > 0.97$ . Picking parameters for Beta distributions that match these statements gives

$$\eta \sim \text{Beta}(19.34, 23.41) \quad \theta \sim \text{Beta}(152.08, 2.53).$$

Two choices for a prior on  $\pi$  are considered: the uniform distribution on  $\pi$  and a Beta(1,5), which is L-shaped with a median of 0.13 and  $\text{Pr}(\pi > 0.1) = (1 - 0.1)^5 = 0.9^5 \doteq 0.6$ . The data consist of a random sample of  $n = 100$  cows. We investigate the posterior distribution of  $\pi$  for various numbers,  $y$ , of cows that test positive and the two priors on  $\pi$ .

Table 14.2 gives the posterior probability of a prevalence below 5%, as well as the posterior median, and a 95% PI for  $\pi$  using the two priors and five values of  $y$ . Differences in results from the two priors are modest.

**EXERCISE 14.8.** The computations in Example 14.2.2 were performed using the following code.

```
model{
  y ~ dbin(PAP, n)
  PAP <- eta*pi + (1-theta)*(1-pi)
```

```

eta ~ dbeta(19.34, 23.41)
theta ~ dbeta(152.08, 2.53)
pi ~ dbeta(1,1)
#pi ~ dbeta(1,5) # Alternate prior
P <- 1-step(pi - 0.05)
}
list(y=0, n=100) # Data
list(eta=0.5, theta=0.5, pi=0.5) # Initial Values

```

Run the code and reproduce the  $y = 5$  results from Table 14.2.

### Gibbs Sampling

A very useful trick in some computer simulations is to sample variables that are actually unobservable. The NGS case provides a particularly nice example. Incorporating simulations of the unobservable  $z_i$ s reduces the computational problem to that of our earlier discussion when a gold standard exists, so the posterior distribution consists of independent Beta distributions. As discussed earlier,

$$p(\pi, \eta, \theta | y, z_1, z_2) = p(\pi | y, z_1, z_2) p(\eta | y, z_1, z_2) p(\theta | y, z_1, z_2).$$

Since the parameters are independent given  $y, z_1, z_2$ , the full conditional distributions needed to perform Gibbs sampling are

$$\begin{aligned} p(\pi | \eta, \theta, y, z_1, z_2) &= p(\pi | y, z_1, z_2) \\ p(\eta | \pi, \theta, y, z_1, z_2) &= p(\eta | y, z_1, z_2) \\ p(\theta | \pi, \eta, y, z_1, z_2) &= p(\theta | y, z_1, z_2). \end{aligned}$$

These are the posterior Beta distributions we gave in (2).

Since we do not actually see  $z_1$  and  $z_2$ , we also need their distributions given the other variables. Given  $\pi$ ,  $\eta$ ,  $\theta$ , and  $y$ , the unobservable random variables  $z_1$  and  $z_2$  are independent with

$$\begin{aligned} z_1 | \pi, \eta, \theta, y &\sim \text{Bin}(y, \text{PVP}) \\ z_2 | \pi, \eta, \theta, y &\sim \text{Bin}(n - y, \text{PVN}). \end{aligned}$$

To implement the Gibbs sampler, we specify starting values  $(\pi^{(0)}, \eta^{(0)}, \theta^{(0)})$ . These determine  $\text{PVP}^{(0)}$  and  $\text{PVN}^{(0)}$ , and are used to sample

$$z_1 \sim \text{Bin}(y, \text{PVP}^{(0)}) \quad z_2 \sim \text{Bin}(n - y, \text{PVN}^{(0)})$$

giving  $(z_1^{(1)}, z_2^{(1)})$ . From the posterior, we then sample parameters given the rest of the parameters

$$\begin{aligned} \pi^{(1)} | \text{rest} &\sim \text{Beta}(z_1^{(1)} + n - y - z_2^{(1)} + a_\pi, y - z_1^{(1)} + z_2^{(1)} + b_\pi), \\ \eta^{(1)} | \text{rest} &\sim \text{Beta}(z_1^{(1)} + a_\eta, n - y - z_2^{(1)} + b_\eta), \\ \theta^{(1)} | \text{rest} &\sim \text{Beta}(z_2^{(1)} + a_\theta, y - z_1^{(1)} + b_\theta). \end{aligned}$$

Using these samples, replace the starting values by  $(\pi^{(1)}, \eta^{(1)}, \theta^{(1)})$  and repeat the process.

One advantage of this scheme is that it is possible to make inferences about the missing observations  $z_1, z_2$ , since approximations to the posteriors for these are generated automatically.

If one does not use the  $z$ s, a different Markov chain simulation method is needed. However, the stationary distributions of both chains, when restricted to only  $(\pi, \eta, \theta)$ , are the same and inferences will be identical up to Monte Carlo error. For Gibbs sampling on only  $(\pi, \eta, \theta)$ , the full conditionals

are different because they are no longer conditional on  $z_1$  and  $z_2$ . In fact, none of them are standard recognizable distributions, so other sampling techniques are needed.

**EXERCISE 14.9.** Write WinBUGS code to handle the NGS problem for the CAD data, assuming that  $z_1$  and  $z_2$  are missing. In doing so, place informative priors on  $\eta$  and  $\theta$  with modes of 0.85 and 0.70, respectively, and 5th percentiles of 0.80 and 0.65, respectively. Assume that this is real prior information. Use a uniform prior for the prevalence. Write code that directly models  $z_1$  and  $z_2$  as appropriate binomials. Another way to handle this is to take the code above and simply put NA in the data list for  $z_1$  and  $z_2$ . WinBUGS automatically knows to get the full conditionals for them. You should also write WinBUGS code without using the  $z$ s. This involves obtaining the likelihood function based on the data  $y$  only, in conjunction with the given priors. Compare results. You should get the same answers using all three of these approaches, but they need not be the same as results presented based on the GS data. See how different they are.

### 14.3 Two Tests, Two Populations

Studies may be conducted to evaluate the performance of one diagnostic test relative to another. For instance, the goal may be to demonstrate that a newly developed test is equivalent or superior to a standard test. The four commonly used study designs for this purpose produce data that are paired (both tests performed on each sampled individual) or unpaired and where a gold standard is available or not. We focus on methods for the most complex setting: paired data with no gold standard. Procedures for the simpler designs are left as exercises.

We will only examine methods for tests that are conditionally independent given the individual; however, this assumption does not always hold. When individuals are tested using more than one diagnostic test, the test outcomes can be dependent (correlated) within the diseased or non-diseased populations. Dependence arises when diagnostic procedures measure similar biological processes, for example, two testing schemes both based on antibodies or two tests based on an antigen.

#### 14.3.1 Methods for Conditionally Independent Tests

A solution to the problem of not being able to identify  $\eta$ ,  $\theta$ , and  $\pi$  in the NGS one test, one population scenario is to make the problem “harder.” We expand the study design from one to two tests, and from one to two populations. This solution was developed by Hui and Walter (1980) who derived maximum likelihood methods. Johnson et al. (2001) detail its Bayesian counterpart. The idea of the study design is that the “first” test is a new test and the second test represents the standard test. The parameters of the model are sensitivities,  $\eta_1$  and  $\eta_2$ , and specificities,  $\theta_1$  and  $\theta_2$ , for each test all of which are common to both populations, and two prevalences,  $\pi_1$  and  $\pi_2$ , one for each population.

In Section 2 we had one test and one population that determined a  $2 \times 2$  table of counts. Now, individuals are sampled from two populations, each has some disease status, and two binary diagnostic tests are applied to each individual so we have a four-dimensional  $2 \times 2 \times 2 \times 2$  table of counts. The factors involved are test 1: + or −, test 2: + or −, population: 1 or 2, and true disease status:  $D$  or  $\bar{D}$ . Unfortunately, with NGS data we do not get to see this four-dimensional table because we do not get to see the disease status. We only observe the three-dimensional table that is collapsed over disease status. Nonetheless, we will need to think about and work with the four-dimensional table.

The observed data, collapsed over disease status, constitute a  $2 \times 2 \times 2$  table of cross-classified diagnostic outcomes as depicted in Table 14.3. In Table 14.3,  $y_{11k}$  is the number of sampled individuals from population  $k$  who tested positive on both tests with  $y_{22k}$  the number who tested negative on both. Similarly,  $y_{12k}$  is the number of sampled individuals from population  $k$  who tested positive on test 1 and negative on test 2 with  $y_{21k}$  the number who were negative on the first and positive on the second. The values  $z_{ijk}$  are used to denote the number of diseased individuals in sample  $k$

Table 14.3: Observed counts  $y_{ijk}$  and missing diseased counts  $z_{ijk}$ .

		Population 1		Population 2	
		$T^+$	$T^-$	$T^+$	$T^-$
Test 1	$T^+$	$y_{111}(z_{111})$	$y_{121}(z_{121})$	$y_{112}(z_{112})$	$y_{122}(z_{122})$
	$T^-$	$y_{211}(z_{211})$	$y_{221}(z_{221})$	$y_{212}(z_{212})$	$y_{222}(z_{222})$

who have test results  $i$  and  $j$ . It follows that  $y_{ijk} - z_{ijk}$  is the corresponding number of non-diseased individuals.

For  $k = 1, 2$ , define a  $2 \times 2$  table (matrix)

$$y_k = \begin{bmatrix} y_{11k} & y_{12k} \\ y_{21k} & y_{22k} \end{bmatrix}.$$

It will be convenient to think of  $y_k$  as both a matrix and as the vector  $[y_{11k}, y_{21k}, y_{12k}, y_{22k}]'$  (which is the Vec operator applied to the matrix  $y_k$ ). Fundamentally, this is just a collection of four numbers identified by the  $ij$  indices (a tensor). Rather than being formal about using the Vec operator, we rely on the context and only identify a specific matrix or vector form when the context is ambiguous. Defining probabilities  $p_k$  similar to  $y_k$ , we have

$$y_k | p_k \sim \text{Mult}(n_k, p_k) \quad p_k = \{p_{ijk} : i, j = 1, 2\}.$$

Here  $p_{11k} = \Pr(T_1^+, T_2^+ | \text{population } k)$  and similar definitions hold for the other three components of  $p_k$ .

Observe that for each population  $k$ , the total count for each table is fixed, so there are 3 degrees of freedom for each  $2 \times 2$  table and thus a total of 6 degrees of freedom for the combined  $2 \times 2 \times 2$  table. We also have a total of 6 parameters:  $\eta_h, \theta_h, h = 1, 2$  and  $\pi_k, k = 1, 2$ . While there is no guarantee that the model will be identifiable just because the number of degrees of freedom (df) is at least as large as the number of parameters, the model we pose here is essentially identifiable (Johnson and Hanson, 2005; Jones et al., 2010). We clarify the meaning of “essentially” in Exercise 14.17. On the other hand, if the number of parameters is larger than the number of df, the model will definitely not be identifiable. If we had only a single population with two tests, we would have 5 parameters but only 3 df, so this model would definitely not be identifiable.

The analysis requires samples from two populations with different prevalences. If we have two populations with the same prevalence then, as seen later,  $p_1 = p_2$  and we are essentially back to a single population, say,  $y_1 + y_2 | p_1 \sim \text{Mult}(n_1 + n_2, p_1)$ . If the available data are a sample from one population, it is tempting to split the data into samples from two subpopulations. An appropriate split of the data is based on some factor, like gender, or a treatment of some kind, where it is expected that prevalences will be different between groups but that tests will behave similarly in both groups.

Our analysis depends on three key model assumptions. First, that the two test outcomes for a given individual are independent conditional on disease status. Second, that the sensitivities and specificities of the two tests remain unchanged across populations. Finally, that the two populations have distinct prevalences. Two tests are conditionally independent if, conditional on the disease status of the individual, the probability of testing positive (or negative) on test 1 is independent of testing positive (or negative) on test 2, e.g.,  $P(T_1^+ | T_2^+, D) = P(T_1^+ | D) = \eta_1$ . Under these conditions, one can show that the multinomial cell probabilities for population  $k$  are given by

$$\begin{aligned} p_{11k} &= \pi_k \eta_1 \eta_2 + (1 - \pi_k)(1 - \theta_1)(1 - \theta_2) \\ p_{12k} &= \pi_k \eta_1 (1 - \eta_2) + (1 - \pi_k)(1 - \theta_1)\theta_2 \\ p_{21k} &= \pi_k (1 - \eta_1)\eta_2 + (1 - \pi_k)\theta_1(1 - \theta_2) \\ p_{22k} &= \pi_k (1 - \eta_1)(1 - \eta_2) + (1 - \pi_k)\theta_1\theta_2. \end{aligned}$$

For example,  $p_{11k}$  is

$$\begin{aligned} p_{11k} &= P(T_1^+, T_2^+ | \text{population } k) \\ &= P(T_1^+, T_2^+ | D)P(D | \text{pop. } k) + P(T_1^+, T_2^+ | \bar{D})P(\bar{D} | \text{pop. } k) \\ &= P(T_1^+ | D)P(T_2^+ | D)P(D | \text{pop. } k) + P(T_1^+ | \bar{D})P(T_2^+ | \bar{D})P(\bar{D} | \text{pop. } k) \\ &= \pi_k \eta_1 \eta_2 + (1 - \pi_k)(1 - \theta_1)(1 - \theta_2) \end{aligned}$$

where the second equality uses the fact that test behavior is the same in each population and the third line uses conditional independence of the tests.

Substantive prior information is typically available for the prevalences and the sensitivity and specificity of the standard test. We assume independent Beta priors for all parameters:

$$\begin{aligned} \pi_1 &\sim \text{Beta}(a_{\pi_1}, b_{\pi_1}), \quad \eta_1 \sim \text{Beta}(a_{\eta_1}, b_{\eta_1}), \quad \theta_1 \sim \text{Beta}(a_{\theta_1}, b_{\theta_1}) \\ \pi_2 &\sim \text{Beta}(a_{\pi_2}, b_{\pi_2}), \quad \eta_2 \sim \text{Beta}(a_{\eta_2}, b_{\eta_2}), \quad \theta_2 \sim \text{Beta}(a_{\theta_2}, b_{\theta_2}). \end{aligned}$$

For a discussion of the relative merits of adding populations to the two-test problem see Gustafson (2005).

**EXAMPLE 14.3.1.** *Nucleospora salmonis* in Trout. Two diagnostic tests for *Nucleospora salmonis* were evaluated using rainbow trout sampled from a California fish hatchery, see Georgiadis et al. (1998) or Enøe et al. (2000). The two tests are (1) microscopic examination (ME) of kidney imprints and (2) a nested polymerase chain reaction (PCR) molecular based assay. Here, the two populations refer to different dates of sampling that occurred approximately 1 year apart. The biological bases for the two tests are different; ME-positive trout are those for which the parasite is visually evident and the PCR test is based on DNA detection, so the two tests are assumed conditionally independent.

The data for the  $n_1 = 132$  trout sampled from population 1 are

$$y_1 = \begin{bmatrix} 0 & 0 \\ 3 & 129 \end{bmatrix}$$

and the  $n_2 = 30$  observations from population 2 are

$$y_2 = \begin{bmatrix} 3 & 0 \\ 24 & 3 \end{bmatrix}.$$

Priors were taken from Enøe et al. (2000) based on information elicited from an expert. The prior mode of  $\theta_1$  was 0.98 and the expert was 95% sure that the specificity of ME was  $> 0.80$ . The elicited mode for the sensitivity of ME was 0.55 with 95th percentile of 0.85. The Beta prior distributions used for  $\eta_2$  and  $\theta_2$  had modal values of 0.90 and 0.85, respectively, with each having 5th percentile equal to 0.60. A Beta prior distribution with mode of 0.30 and 5th percentile of 0.08 was used for  $\pi_2$ , the infection prevalence for population 2, and for  $\pi_1$  the mode was 0.03 with 95th percentile 0.30.

The posterior medians with 95% intervals for test accuracy presented in Table 14.4 are based on fitting the model with the informative priors and with  $U(0, 1)$  reference priors used for all parameters. The ME test appears to be a disaster. Not only is the sensitivity shockingly low, but it is far below prior expectations. The posterior 95th percentiles are well below the informative prior mode. The data are also quite unusual. In population 1, the tests agree 98% of the time but in population 2 they disagree 80% of the time. The PCR test appears to be superior because it has a substantially higher sensitivity with comparable specificity.

**EXERCISE 14.10.** WinBUGS code for analyzing the *Nucleospora salmonis* data without modeling the latent  $z$ 's is given below.

Table 14.4: Posterior medians and 95% PIs of test accuracy parameters for detecting N. salmonis in trout.

	$\eta_1$	$\theta_1$	$\eta_2$	$\theta_2$
Informative priors	0.166 (0.066, 0.316)	0.993 (0.972, 0.999)	0.938 (0.815, 0.992)	0.967 (0.928, 0.991)
Uniform priors	0.118 (0.037, 0.258)	0.995 (0.972, 1.0)	0.935 (0.779, 0.997)	0.983 (0.945, 0.999)

```

model
{
  y1[1:2, 1:2] ~ dmulti(p1[1:2, 1:2], n1)
  y2[1:2, 1:2] ~ dmulti(p2[1:2, 1:2], n2)
  p1[1,1] <- pi1*eta1*eta2 + (1-pi1)*(1-theta1)*(1-theta2)
  p1[1,2] <- pi1*eta1*(1-eta2) + (1-pi1)*(1-theta1)*theta2
  p1[2,1] <- pi1*(1-eta1)*eta2 + (1-pi1)*theta1*(1-theta2)
  p1[2,2] <- pi1*(1-eta1)*(1-eta2) + (1-pi1)*theta1*theta2
  p2[1,1] <- pi2*eta1*eta2 + (1-pi2)*(1-theta1)*(1-theta2)
  p2[1,2] <- pi2*eta1*(1-eta2) + (1-pi2)*(1-theta1)*theta2
  p2[2,1] <- pi2*(1-eta1)*eta2 + (1-pi2)*theta1*(1-theta2)
  p2[2,2] <- pi2*(1-eta1)*(1-eta2) + (1-pi2)*theta1*theta2
  eta1 ~ dbeta(2.82, 2.49)
  theta1 ~ dbeta(15.7, 1.30)
  eta2 ~ dbeta(8.29, 1.81)
  theta2 ~ dbeta(10.69, 2.71)
  pi1 ~ dbeta(1.27, 9.65)
  pi2 ~ dbeta(1.73, 2.71)
}
list(n2=30,n1=132,y1=structure(.Data=c(0,0,3,129),.Dim=c(2,2)),
     y2=structure(.Data=c(3,0,24,3), .Dim=c(2,2)))
list(pi1=0.03,pi2=0.30,eta1=0.55,
     theta1=0.98,eta2=0.90,theta2=0.85)

```

Reproduce Table 14.4 and extend the code to get estimates of PVP, PVN, and the DLRs. Evaluate the hypothesis that the PCR test is uniformly superior to ME, i.e., test  $H_0 : \eta_2 > \eta_1, \theta_2 > \theta_1$ . Also, obtain  $\Pr(\eta_2 > \eta_1 | y_1, y_2)$  and  $\Pr(\theta_2 > \theta_1 | y_1, y_2)$ . Comment.

The analysis was based on the assumption of equal test accuracy at the two sampling dates. The assumption of constant sensitivity may be questionable because the disease was just beginning (endemic and subclinical) at the first sampling but was evident at the second sampling. Hence, it might be expected that the sensitivity would be higher at the second sampling date. This assumption can be checked by conducting separate analyses for each sampled population, as was done in Enøe et al. (2000). For example, in a particular illustration Enøe et al. (2000) conducted separate analyses and found no appreciable differences from the joint analysis results. These separate analyses would involve tables with only 3 df each and so the prior distributions play a stronger role in the joint analysis. The priors naturally pull the test accuracy estimates toward the prior guesses, which are the same in both populations, but more than the priors, this is probably an artifact of the data. In  $y_1$  there is low prevalence, so most of the available information is about specificity, which looks good. In  $y_2$  there is high prevalence, so the information is about sensitivity and the tests behave very differently. Moreover, there are far more data in  $y_1$ , so the  $\theta_h$  intervals are much narrower. Because  $y_2$  has fewer observations, the  $\eta_h$  intervals are wider.

**EXERCISE 14.11.** (a) Analyze the trout data separately, obtaining two analyses that correspond to the two populations, all using the same priors as above. To do this, you must revise the code to handle a single  $2 \times 2$  table, and then just run the code twice with the two different data sets. (b) Revise the code so that both samples can be analyzed at once, except now under the assumption that sensitivities and specificities are different across the two populations. In this analysis, you should make inferences about the difference in sensitivities across populations, and the difference in specificities across populations. You will now need independent priors for four sensitivities and four specificities.

#### Posterior Calculations\*

As in the single-test setting, the most elegant way to sample the posterior distribution augments the observed counts with latent counts of diseased individuals in each category (see Table 14.3). The introduction of latent data  $z_k = [z_{11k}, z_{21k}, z_{12k}, z_{22k}]'$  facilitates posterior simulation via Gibbs sampling since the full conditional distributions are all recognizable. We now discuss this but note that the introduction of the  $z_k$ s is not necessary in WinBUGS, since in their absence WinBUGS will instead sample the unrecognizable full conditionals using Metropolis steps within the Gibbs sampler (this follows since the *adapting* box in the *Update Tool* is checked in WinBUGS).

The augmented data likelihood is determined by the joint distribution of  $(y_1, z_1)$  and  $(y_2, z_2)$ . Suppressing parameters,  $f(y_k, z_k) = f(z_k|y_k)f(y_k)$  where  $f(y_k)$  is the multinomial probability function corresponding to  $y_k \sim \text{Mult}(n_k, p_k)$  and  $f(z_k|y_k)$  is the product of four independent binomials, one corresponding to each cell in the table for  $y_k$ . For example,

$$z_{11k}|y_{11k} \sim \text{Bin}\left(y_{11k}, P(D|T_1^+, T_2^+, \text{population } k)\right).$$

Using Bayes' Theorem, the fact that tests are conditionally independent, and that test performance does not depend on the population

$$\begin{aligned} P(D|T_1^+, T_2^+, \text{population } k) &= \frac{P(T_1^+, T_2^+|D)P(D|\text{pop. } k)}{P(T_1^+, T_2^+|\text{pop. } k)} \\ &= \frac{P(T_1^+|D)P(T_2^+|D)P(D|\text{pop. } k)}{p_{11k}} \\ &= \frac{\eta_1 \eta_2 \pi_k}{p_{11k}}. \end{aligned}$$

Similar derivations yield

$$\begin{aligned} z_{12k}|y_{12k} &\sim \text{Bin}\left(y_{12k}, \frac{\eta_1(1-\eta_2)\pi_k}{p_{12k}}\right) \\ z_{21k}|y_{21k} &\sim \text{Bin}\left(y_{21k}, \frac{(1-\eta_1)\eta_2\pi_k}{p_{21k}}\right) \\ z_{22k}|y_{22k} &\sim \text{Bin}\left(y_{22k}, \frac{(1-\eta_1)(1-\eta_2)\pi_k}{p_{22k}}\right). \end{aligned}$$

Putting the multinomial and binomial components together gives the augmented data likelihood based on  $y = (y_1, y_2)$  and  $z = (z_1, z_2)$

$$\begin{aligned} L(\pi_1, \pi_2, \eta_1, \eta_2, \theta_1, \theta_2 | y, z) &\propto \prod_{k=1}^2 p_{11k}^{y_{11k}} p_{12k}^{y_{12k}} p_{21k}^{y_{21k}} p_{22k}^{y_{22k}} \\ &\quad \times \left[ \frac{\eta_1 \eta_2 \pi_k}{p_{11k}} \right]^{z_{11k}} \left[ \frac{(1-\theta_1)(1-\theta_2)(1-\pi_k)}{p_{11k}} \right]^{y_{11k}-z_{11k}} \end{aligned}$$

$$\begin{aligned}
& \times \left[ \frac{\eta_1(1-\eta_2)\pi_k}{p_{12k}} \right]^{z_{12k}} \left[ \frac{(1-\theta_1)\theta_2(1-\pi_k)}{p_{12k}} \right]^{y_{12k}-z_{12k}} \\
& \times \left[ \frac{(1-\eta_1)\eta_2\pi_k}{p_{21k}} \right]^{z_{21k}} \left[ \frac{\theta_1(1-\theta_2)(1-\pi_k)}{p_{21k}} \right]^{y_{21k}-z_{21k}} \\
& \times \left[ \frac{(1-\eta_1)(1-\eta_2)\pi_k}{p_{22k}} \right]^{z_{22k}} \left[ \frac{\theta_1\theta_2(1-\pi_k)}{p_{22k}} \right]^{y_{22k}-z_{22k}} \\
= & \eta_1^{z_{1..}} (1-\eta_1)^{z_{2..}} \eta_2^{z_{..1}} (1-\eta_2)^{z_{..2}} \\
& \times \theta_1^{y_{1..}-z_{1..}} (1-\theta_1)^{y_{1..}-z_{1..}} \theta_2^{y_{2..}-z_{2..}} (1-\theta_2)^{y_{2..}-z_{2..}} \\
& \times \pi_1^{z_{..1}} (1-\pi_1)^{n_1-z_{..1}} \pi_2^{z_{..2}} (1-\pi_2)^{n_2-z_{..2}}.
\end{aligned}$$

With independent Beta prior distributions, the full conditionals for the model parameters are Beta distributions:

$$\begin{aligned}
\eta_1 | \text{rest} &\sim \text{Beta}(a_{\eta_1} + z_{1..}, b_{\eta_1} + z_{2..}) \\
\eta_2 | \text{rest} &\sim \text{Beta}(a_{\eta_2} + z_{..1}, b_{\eta_2} + z_{..2}) \\
\theta_1 | \text{rest} &\sim \text{Beta}(a_{\theta_1} + y_{1..} - z_{1..}, b_{\theta_1} + y_{2..} - z_{1..}) \\
\theta_2 | \text{rest} &\sim \text{Beta}(a_{\theta_2} + y_{..2} - z_{..2}, b_{\theta_2} + y_{..1} - z_{..1}) \\
\pi_1 | \text{rest} &\sim \text{Beta}(a_{\pi_1} + z_{..1}, b_{\pi_1} + n_1 - z_{..1}) \\
\pi_2 | \text{rest} &\sim \text{Beta}(a_{\pi_2} + z_{..2}, b_{\pi_2} + n_2 - z_{..2}).
\end{aligned}$$

The binomials for  $z|y$  listed earlier are used for simulating the missing  $z_k$ s. Given starting values for the parameters, Gibbs sampling proceeds by successively simulating from these distributions, substituting current values at each step. Inferences comparing test accuracies (and for prevalences, PVP, PVN, and DLRs) are easy to make using the Gibbs iterates as illustrated in Exercise 14.12.

**EXERCISE 14.12.** Revise the code in Exercise 14.10 to use the  $z$ s for making inferences for the trout data. Compare inferences with those obtained in Exercise 14.10. In addition, make inferences for the  $z$ s, and perform a sensitivity analysis. Also, make posterior inferences that will ascertain whether the ME specificity is preferable to that for the PCR, and similarly regarding the sensitivity. Comment on assumptions.

#### 14.4 Prevalence Distributions

We now examine how disease prevalence varies across subgroups (clusters). Doing this involves a two-stage cluster sampling plan. First, a random sample of  $k$  clusters (e.g. villages, cities, herds) is selected followed by a random sample of  $n_i$  individuals from the  $i$ th cluster. Every subject in the survey is tested using one or more diagnostic tests. We are interested in drawing conclusions for the entire population but also in learning about the *prevalence distribution*, i.e., how prevalences vary from cluster to cluster.

**EXAMPLE 14.4.1.** *National Prevalence of Ovine Progressive Pneumonia.* Branscum et al. (2005) examined results from a serological survey of ovine progressive pneumonia (OPP) in U.S. sheep. The survey was conducted by the National Animal Health Monitoring System (NAHMS) in 2001. It sampled flocks from across the country. In each flock, serum samples were collected from between 3 and 40 sheep (mean of 31 samples per flock) and tested using an ELISA test for OPP. We consider data from 319 flocks that had at least one seropositive sheep.

We consider only (presumably) infected flocks to keep the mathematics of the example simple. If there is a substantial number of uninfected flocks, the prevalence distribution should put positive probability on prevalences of 0. We want a continuous model for the prevalence distribution, so we

consider only infected flocks. Unfortunately, without gold standard data available, we do not know which flocks are infected, so we treat the presumably infected flocks (those with positive tests) as a sample of infected flocks.

Throughout, we assume that sensitivities and specificities remain constant over clusters as in Section 3. For simplicity, we consider surveys that involve giving only one test to each individual. With a single test having sensitivity  $\eta$ , specificity  $\theta$ , and a cluster with prevalence  $\pi_i$ , the cluster's "population apparent prevalence" is  $p_i = \eta\pi_i + (1 - \theta)(1 - \pi_i)$ . The test data are binomial samples taken from each of the  $k$  clusters,

$$y_i \sim \text{Bin}(n_i, p_i), \quad p_i = \eta\pi_i + (1 - \theta)(1 - \pi_i), \quad i = 1, \dots, k.$$

Typically, we aren't able to say which clusters would have higher or lower prevalences, so the prevalences are considered exchangeable (iid conditional on some parameters). We treat the cluster prevalences  $\pi_i$  as a random sample from a distribution of prevalences. Our main goal is to estimate this *prevalence distribution*. Assume

$$\pi_1, \pi_2, \dots, \pi_k | \mu, \psi \stackrel{iid}{\sim} \text{Beta}(\mu\psi, (1 - \mu)\psi), \quad (1)$$

so that  $E(\pi_i) = \mu$ , and  $\text{Var}(\pi_i | \mu, \psi) = \mu(1 - \mu)/(\psi + 1)$ . A large value of  $\psi$  implies less variability among prevalences. This model for prevalences has its shortcomings. If there were two different types of clusters, one with higher prevalences than the other, we might expect a bimodal prevalence distribution which a Beta distribution cannot model, although a mixture of two Betas could. We proceed using a single Beta distribution.

As prior distributions for the parameters of the prevalence distribution we assume

$$\mu \sim \text{Beta}(a_\mu, b_\mu) \quad \perp \quad \psi \sim \text{Gamma}(a_\psi, b_\psi).$$

To compute the marginal and posterior prevalence distributions, let  $B(\pi | \mu, \psi)$  denote the Beta density corresponding to (1), and let  $p_1(\mu)$  and  $p_2(\psi)$  denote the prior densities for  $\mu$  and  $\psi$ . Then define the prior and (posterior) predictive prevalence densities to be

$$\begin{aligned} p_*(\pi) &= \int B(\pi | \mu, \psi) p_1(\mu) p_2(\psi) d\mu d\psi \\ p_*(\pi | \text{data}) &= \int B(\pi | \mu, \psi) p(\mu, \psi | \text{data}) d\mu d\psi. \end{aligned} \quad (2)$$

In either case, the prevalence density is a continuous mixture of Beta densities. In practice, we cannot compute the prevalence density for every  $\pi$  value between 0 and 1. We must evaluate the density over a fine partition of the unit interval. We also approximate the integrals in (2) using a discrete approximation to the joint distribution of  $\mu$  and  $\psi$ .

To obtain prior information about  $\mu$ , as usual we elicit an expert's best guess  $\mu_0$  for the mean  $\mu$  of the prevalence distribution and a percentile for  $\mu$ . The expert needs to be alerted that this is a percentile for the mean and not a percentile for the prevalence distribution. For example, if  $\mu_0 = 0.3$  and the 95th percentile is 0.4, the prior for  $\mu$  is Beta(20.9, 47.5).

Information about  $\psi$  is more difficult. We follow our basic principle of eliciting information about things that a scientist can think about and then inducing a distribution on  $\psi$ . The first step is to obtain a best guess for  $\psi$ , say  $\psi_0$ . By assumption, the prevalence distribution is Beta( $\mu\psi, (1 - \mu)\psi$ ). We are going to take a best guess for the prevalence distribution, say, Beta( $a_0, b_0$ ) and equate it to Beta( $\mu_0\psi_0, (1 - \mu_0)\psi_0$ ). Then, regardless of the value of  $\mu_0$ , we have  $\psi_0 = a_0 + b_0$ .

To obtain a prior guess for the prevalence distribution, we already have a prior guess for the mean  $\mu$ , which is  $\mu_0$ . Additionally, we ask our expert to think about the value that is exceeded by only, say, 10% of the prevalences in the population of prevalences. Let  $\gamma$  be this 90th percentile and let  $\gamma_0$  be the scientist's best guess for the 90th percentile of the prevalence distribution. We now find the Beta

distribution with mean  $\mu_0$  and 90th percentile  $\gamma_0$ , our  $\text{Beta}(a_0, b_0)$ . For example, if our best guess for the mean of the prevalence distribution is  $\mu_0 = 0.3$  and our best guess for the 90th percentile of the prevalence distribution is  $\gamma_0 = 0.5$ , we need to find the  $\text{Beta}(a_0, b_0)$  that has  $0.3 = a_0/(a_0 + b_0)$  and has 0.5 as the 90th percentile. (Using BetaBuster, this requires some trial and error.) The distribution turns out to be  $\text{Beta}(2.85, 6.54)$ , so our best guess for  $\psi$  is  $\psi_0 = 2.85 + 6.54 = 9.39$ .

By assumption,  $\psi \sim \text{Gamma}(a_\psi, b_\psi)$ , so we need to determine values for  $a_\psi$  and  $b_\psi$ . First take the best guess  $\psi_0$  and equate it to the mode of the gamma distribution, i.e., set  $\psi_0 = (a_\psi - 1)/b_\psi$  or equivalently,  $a_\psi = 1 + b_\psi \psi_0$ . To determine both values  $a_\psi$  and  $b_\psi$ , we need another equation. We have found it difficult to obtain from scientists the nuanced input necessary to proceed. Hanson et al. (2003a) discuss alternative methods. For simplicity, we here fall back on the idea of reference priors. We select  $b_\psi$  to be small so as to create a large variance for the prior on  $\psi$ . With a Gamma distribution, the standard deviation becomes  $\sqrt{a_\psi/b_\psi^2} = \sqrt{(1+b_\psi\psi_0)/b_\psi}$ . Selecting  $b_\psi = 0.01$  results in a standard deviation of 105, and  $b_\psi = 0.1$  results in a standard deviation of 14 for this example. Since our best guess for  $\psi$  was 9.39, we see that two standard deviations above the mean is about 37 in the latter case, which is quite far away from our guess, so we would pick  $b_\psi = 0.1$  in this instance. In performing a data analysis, we recommend trying different values of  $b_\psi$  to see how sensitive posterior results are to this choice.

**EXERCISE 14.13.** Using these priors on  $\mu$  and  $\psi$ , write WinBUGS code to obtain the prior prevalence density over a grid of  $\pi$  values. Give the density for 10 equally spaced points in your grid, or better, give a plot over your grid. Also obtain numerical approximations for the prior probability that a randomly selected cluster will have a prevalence that is larger than 0.2, 0.4, 0.5, and 0.6, respectively.

**EXAMPLE 14.4.1 CONTINUED. *Prevalence of Ovine Progressive Pneumonia.*** The NAHMS survey of OPP had 319 flocks with at least one seropositive sheep. The mean of the apparent prevalences among these flocks was 0.32. The raw data are subject to confidentiality agreements between NAHMS and study participants, but we present a posterior analysis.

Prior information comes from a validation study, Marshall (2003). The sensitivity  $\eta$  and specificity  $\theta$  of the ELISA test had 95% prior intervals of (0.90, 0.99) and (0.77, 0.97), respectively. The 95% prior interval for the mean  $\mu$  of the prevalence distribution was (0.05, 0.55) and our best guess of the 90th percentile of the prevalence distribution is  $\gamma_0 = 0.34$ . Assuming Beta distributions, these intervals determine their two parameters. In particular,  $\eta \sim \text{Beta}(71, 3)$ ,  $\theta \sim \text{Beta}(32, 4)$ , and  $\mu \sim \text{Beta}(2.64, 7.55)$ . Translating these into our usual criteria, the expert's best guess for  $\eta$  was 0.97 and they were 95% sure that the sensitivity was at least 0.92. Their best guess of  $\theta$  was 0.91 and they were 99% sure that the specificity was at least 0.74. Their best guess for  $\mu$  was  $\mu_0 = 0.20$  and they were 95% sure that the mean prevalence was less than 0.50. The  $\text{Beta}(a_0, b_0)$  distribution with a mean of 0.20 and 90th percentile of 0.34 corresponds to  $\mu_0 = 0.20$ ,  $\psi_0 = 15$ , so we set  $a_\psi = 1 + 15b_\psi$ .  $b_\psi$  was selected to be 0.20, which results in a prior standard deviation for  $\psi$  of 10. The prior based predictive prevalence density is given in Figure 14.1.

The mean flock prevalence  $\mu$  has posterior median 0.28 with a 95% probability interval of (0.25, 0.32). The prior median and interval were 0.24 and (0.05, 0.55), so we have much more certainty about  $\mu$  after incorporating the data. However, we have more spread in the predictive (or posterior) prevalence distribution than in the prior prevalence distribution. The prior 80% interval for prevalences was (0.12, 0.50) whereas the posterior interval is (0.003, 0.76). This does not contradict the idea that more data give us better results. More data allow us to know parameters more accurately and learn more about the prevalence distribution. So with more data we should arrive at a consensus. However, there is no reason to think that more data would give us a less variable prevalence distribution. More data merely tell us more accurately what the prevalence distribution is. Prior and posterior prevalence distributions are plotted in Figure 14.1. To see that the data make us more certain about the prevalence distribution, consider the 90th percentile of the conditional (on  $\mu$  and  $\psi$ )

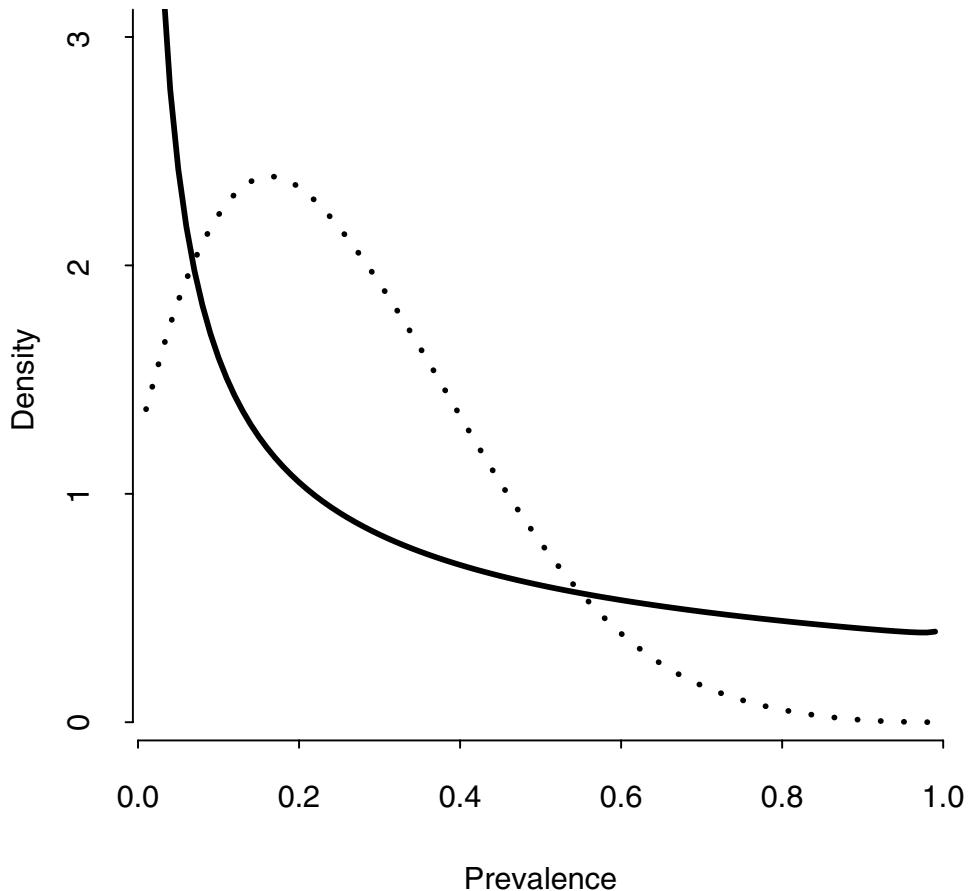


Figure 14.1: Prior (dotted line) and posterior (solid line) prevalence densities for OPP.

prevalence distribution. This has a prior 95% probability interval of (0.12, 0.72), whereas the posterior interval is (0.69, 0.83), which is much narrower. Note that in both the prior and the posterior prevalence distributions, we have integrated out the variability in  $\mu$  and  $\psi$ .

Estimates of diagnostic test sensitivity and specificity should not be a primary goal of this study design; nonetheless, they are available. The posterior median and 95% interval for  $\eta$  are 0.98, (0.96, 0.995), and for  $\theta$  are 0.95, (0.94, 0.97). The corresponding interval estimates from the prior distribution are (0.90, 0.99) for  $\eta$  and (0.77, 0.97) for  $\theta$ . The prior based intervals here were elicited directly from the expert and will not conform exactly to the selected beta priors.

**EXERCISE 14.14.** Consider WinBUGS code to fit the prevalence estimation model based on data from a single test.

```
model{
  for(i in 1:k){
    y[i] ~ dbin(p[i],n[i])
    p[i] <- eta*pi[i] + (1-theta)*(1-pi[i])
    pi[i] ~ dbeta(alpha,beta)
  }
  PPD ~ dbeta(alpha,beta) # Generates
  alpha <- mu*psi          # pred prev dens
```

```

beta <- psi*(1-mu)
mu ~ dbeta(2.64,7.55)
psi ~ dgamma(4,0.2)
eta ~ dbeta(71,3)
theta ~ dbeta(32,4)
}

```

The NAHMS data are not available but find the prior prevalence density by modifying this code to eliminate the data. Observe that the posterior prevalence density is obtained by the simple addition of PPD in the code. Give a convincing argument that this gives the appropriate density as output by monitoring PPD. This eliminates the need to use a grid of values as in the previous exercise.

**EXERCISE 14.15.** Estimate the prevalence distribution for hypopneumonia in swine. Nineteen herds were sampled with various sample sizes taken from each herd. The first herd is *known to be non-infected*. The priors were elicited from an expert. The prior on  $\psi$  was not elicited. Run the code given here, and then modify it to account for there being no infected animals in population 1. Do a sensitivity analysis, especially with regard to the prior on  $\psi$ .

```

model{
  for(i in 1:k){
    y[i] ~ dbin(p[i],n[i])
    p[i] <- pi[i]*eta + (1-pi[i])*(1-theta)
    pi[i] ~ dbeta(alpha,beta)I(0.0001,0.9999)
  }
  alpha <- mu*psi
  beta <- (1-mu)*psi
  PPD ~ dbeta(alpha,beta) # This will generate
                           # the pred prev density
  mu ~ dbeta(1,10)
  psi ~ dgamma(1,1)
  eta ~ dbeta(24,3)
  theta ~ dbeta(38,2)
}
list(pi=c(0.0001,0.7,0.7,0.7,0.7,0.7,0.7,0.7,0.7,
        0.7,0.7,0.7,0.7,0.7,0.7,0.7,0.7,0.7),
     eta=0.9, theta=0.95, mu=0.7, psi=10, PPD=0.7)
list(k=19, n=c(2128,20,20,19,20,12,20,19,29,10,
             21,20,20,21,20,20,21,21,20),
     y=c(5,15,13,4,14,10,17,12,10,9,20,8,16,20,13,6,12,20,17))

```

**EXERCISE 14.16.** We assume a prevalence distribution of  $\text{Beta}(\mu\psi, (1 - \mu)\psi)$ . Increasing  $\psi$  decreases the variance. We seek a way of obtaining a uniform prior on  $\psi$  where we determine lower and upper values, say  $\psi_l$  and  $\psi_u$ , for which we are virtually certain  $\psi \in [\psi_l, \psi_u]$ . Then, to obtain a prior on  $\psi$ , think about the prevalence distribution when  $\mu$  is our best guess, 0.2. Suppose that you are certain that the 99th percentile of the prevalence distribution is no larger than 0.7. Find  $\psi_l$  so that a  $\text{Beta}(0.2\psi, (1 - 0.2)\psi)$  distribution has 99th percentile 0.7. Similarly, suppose you are certain that the 99th percentile must exceed 0.3 and find the value of  $\psi_u$  that gives the beta distribution with that 99th percentile. If we are certain that  $\psi$  must be between these two values, we might place a uniform prior on  $\psi$  values between them, namely  $U[\psi_l, \psi_u]$ . Reanalyze the data of Exercise 14.15 using this prior and compare results.

There are many extensions of the material presented here. For example, as previously mentioned, diagnostic tests are often conditionally dependent, rather than independent. There are often more

than two tests. And it is common to use two or more tests when estimating prevalence distributions. Moreover, diagnostic tests are not always dichotomous, they can often be regarded as continuous. Readers may wish to visit the diagnostic testing website referred to at the beginning of this chapter to see how these problems are handled using WinBUGS. We have also placed a chapter on continuous diagnostic tests, originally intended for this book, in the Chapter 14 repository on the book website.

EXERCISE 14.17. The model with two tests in two populations is weakly nonidentifiable. What that means here is that for every parameter vector  $(\pi_1, \pi_2, \eta_1, \eta_2, \theta_1, \theta_2)$ , there is another unique vector that generates exactly the same probability distribution for the multinomial data. In other words, it is possible to get estimates that are pure nonsense. You could get estimates of sensitivity and specificity that are very small when it is known that the tests are reasonably accurate. The other vector that generates the same model is  $(1 - \pi_1, 1 - \pi_2, 1 - \theta_1, 1 - \theta_2, 1 - \eta_1, 1 - \eta_2)$ . Substituting these values into the formulas just before Example 14.3.1 results in exactly the same  $p_{ijk}$ s and consequently the same model. A simple way of avoiding this problem is to use initial values that are the prior modes. Run the code for the trout data with starting values that are consistent with the wrong answers and compare results.

---

## Chapter 15

---

# Nonparametric Models

---

Measurements on the heights of a random sample of people are likely to be bimodal with women noticeably shorter on average than men. There are two ways to handle such data. If we know each individual's sex, i.e., if the covariate "sex" has been measured, we can examine the women and men separately, using, say, a normal distribution for each group. However, if we do not have the covariate available, we need a more general class of distributions than the normal to deal with the bimodal behavior. This is a very simple example but it conveys the spirit of nonparametric modeling. When the standard models that we have used fail to capture the salient aspects of the data, we need to develop more general models that are appropriate. The most common nonparametric models come in two flavors: incorporating more general families of distributions and incorporating more general mean structures. We can use broader classes of distributions or we can use more complicated regression functions but we seldom generalize both simultaneously.

In Chapters 7 through 11 we dealt with reasonably simple regression problems that involved simple distributions such as the binomial and normal and reasonably simple regression functions. In this chapter we will expand on both aspects of those procedures. Thus far our distributional models for data have been built on families indexed by one or two parameters such as  $\text{Bern}(p)$ ,  $N(\mu, 1/\tau)$ ,  $\text{Pois}(\lambda)$ ,  $\text{Exp}(\lambda)$ , and  $\text{Weib}(\alpha, \beta)$ . We now present some broader classes of distributions that involve many more parameters. Our earlier regression functions were known functions of unknown linear combinations of the predictor variables. Our treatment of nonparametric modeling of regression functions is (theoretically) in the same vein but the applications are considerably more complicated.

Many generalizations of the regression function can be viewed as straightforward adaptations of the procedures illustrated in Chapters 7 through 11. One approach is to use the current predictors to define additional predictors and then, as before, use known functions of unknown linear combinations from this expanded set of predictors to model regression functions. Another approach uses the simpler methods from earlier chapters but on subsets of the data and then combines the information from the subsets. Both approaches involve fitting many more parameters to the data.

Nonparametric models are anything but nonparametric. These models involve many parameters. In our discussion parameters are added to a basic model to increase the possible shapes for either the density function or the regression function. In Section 1 we discuss more general ways to model distributions. In Section 2 we examine more general ways for defining regression functions. In Section 3 we briefly discuss the application of Section 2 to estimating a baseline hazard for the Proportional Hazards model of Section 13.2.

We provide WinBUGS code on our website for many of the procedures illustrated here. Additionally, DPPackage (Jara, 2007; Jara et al., 2009) is an R package providing a library of functions for fitting various Bayesian nonparametric models for density estimation, generalized linear mixed models, generalized additive models, receiver operator characteristic curve analyses, meta-analysis, dependent processes, survival analysis, etc.

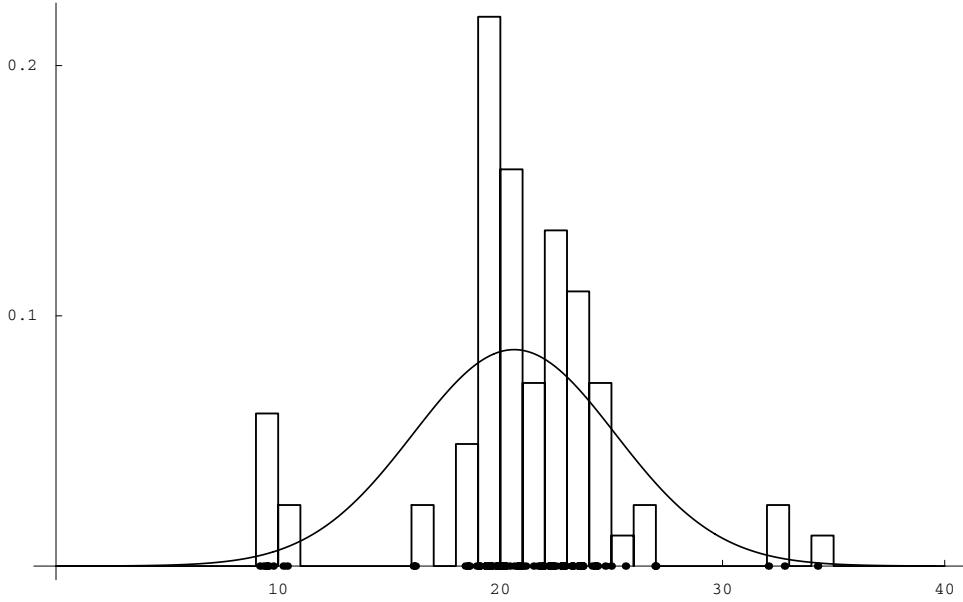


Figure 15.1: Galaxy data: histogram and normal fit.

### 15.1 Flexible Density Shapes

We focus on three methods for defining more flexible distributions: finite mixtures, Dirichlet process mixtures, and mixtures of Polya trees. The following example will be used to illustrate concepts.

**EXAMPLE 15.1.1.** *Galaxy Data.* Figure 15.1 shows a dot plot and histogram of  $n = 82$  galaxy velocities obtained from Roeder (1990). Also on the plot is the predictive density from fitting a normal model with reference priors: specifically

$$\begin{aligned} y_1, \dots, y_{82} | \mu, \tau &\stackrel{iid}{\sim} N(\mu, \tau^{-1}), \\ \mu &\sim N(0, 1000) \quad \perp \!\!\! \perp \quad \tau \sim \text{Gamma}(0.001, 0.001). \end{aligned}$$

The simple normal model does not capture the three well-separated “clumps” of observations seen in the data.

#### 15.1.1 Finite Mixtures

One way to enrich the class of possible density shapes is to consider a finite weighted mixture of two or more parametric distributions, cf. Sections 4.14 and 11.2. For the simple mixture model involving men and women’s heights, let  $y$  be a random height. Let women’s heights be  $X_1 \sim N(\mu_1, 1/\tau_1)$ , men’s heights be  $X_2 \sim N(\mu_2, 1/\tau_2)$ , and  $W$  be an (unobserved) 0-1 indicator of sex with  $W \sim \text{Bern}(p)$ . An observation  $y$  has the distribution of  $y = WX_1 + (1 - W)X_2$ . In other words, with probability  $p$  the observation comes from  $X_1$ , otherwise it comes from  $X_2$ . This mixture model has 5 parameters and density

$$f(y | \mu_1, \mu_2, \tau_1, \tau_2, p) = \frac{p}{\sqrt{2\pi/\tau_1}} e^{-0.5(y-\mu_1)^2\tau_1} + \frac{1-p}{\sqrt{2\pi/\tau_2}} e^{-0.5(y-\mu_2)^2\tau_2}. \quad (1)$$

Figure 15.2 shows some of the variety displayed by the densities in this class. Two of the curves are bimodal, but very different. The third curve is unimodal but skewed.

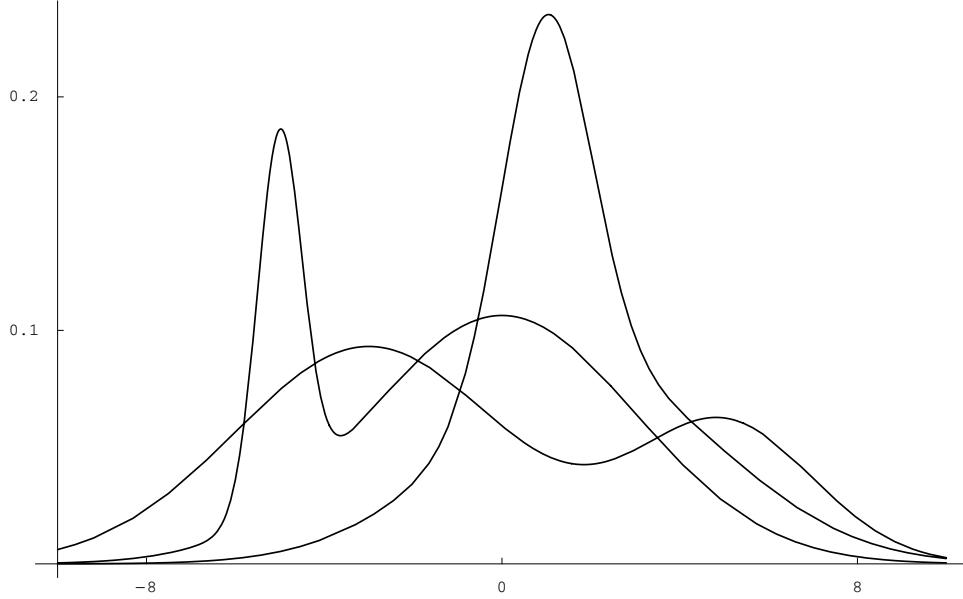


Figure 15.2: Examples of two component normal densities.

Finite mixture models often arise when the overall population is a combination of distinct subpopulations, e.g., sexes, genetic makeups, or subspecies. If the subpopulation for each observation is known, it is probably best to analyze the subpopulations separately and, if necessary, recombine the results for the overall population. Often the subpopulations are unknown and we must work directly with a mixture model.

**EXERCISE 15.1.** Show that the density in (1) integrates to one. Find the cdf in terms of the standard normal cdf  $\Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-0.5z^2} dz$ .

In general, imagine  $K$  subpopulations each characterized with a density  $\phi_k$  involving a vector parameter  $\theta_k$  of dimension  $r_k$  and let  $w_k$  be the proportion of the overall population that comes from the  $k$ th subpopulation. The density is

$$f(y|\theta_1, \dots, \theta_K; w_1, \dots, w_K) = \sum_{k=1}^K w_k \phi_k(y|\theta_k).$$

Most often we take the  $\phi_k$ s all the same, so

$$f(y|\theta_1, \dots, \theta_K; w_1, \dots, w_K) = \sum_{k=1}^K w_k \phi(y|\theta_k)$$

with  $r_1 = \dots = r_K \equiv r$ . For example, in the normal mixture model we take  $\theta_k = (\mu_k, \tau_k)'$  and each  $\phi(y|\mu_k, \tau_k)$  is the density of a  $N(\mu_k, 1/\tau_k)$ .

Placing a prior on  $K$  leads to a model that changes dimension (number of parameters) with  $K$ . Such *trans-dimensional* models are difficult to fit. One approach uses *reversible jump MCMC*. The WinBUGS add-on *Jump* provides this capability. We will not discuss it further.

It is easier to analyze these models if you know (or pretend to know) the finite value of  $K$ . Non-Bayesians typically fit such models by first estimating  $K$ , then estimating the remaining model parameters, typically using maximum likelihood (via the *EM algorithm*). The number of components  $K$  can be estimated using model selection criteria as discussed in Section 4.9. We recommend a cross-validatory measure such as LPML.

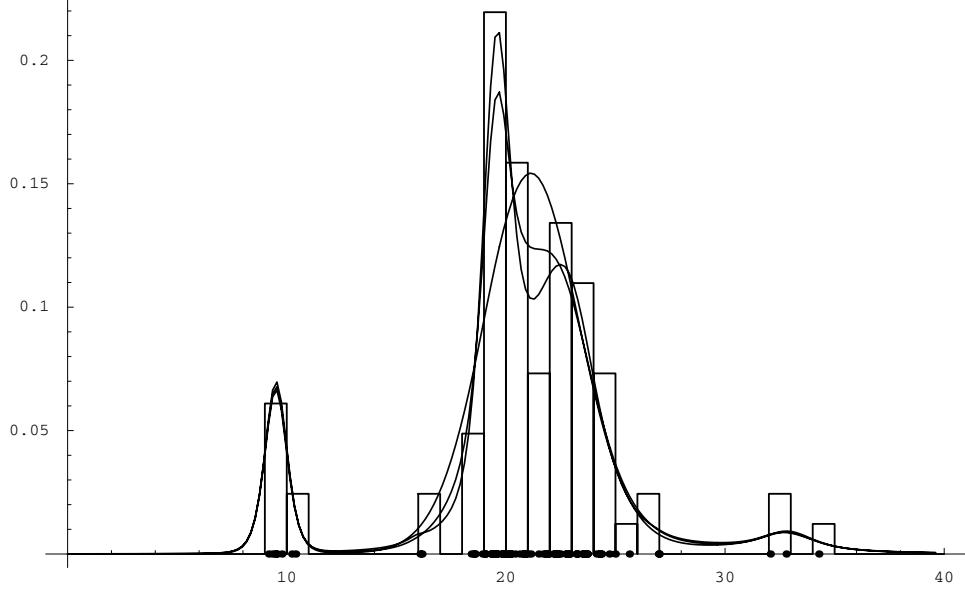


Figure 15.3: Galaxy data: fits from finite mixture models,  $K = 3, 4, 6$ .

**EXAMPLE 15.1.2.** *Galaxy Data.* To use a mixture model on these data, the first question is “How big should  $K$  be?” Looking at Figure 15.1,  $K = 3$  or  $K = 4$  might suffice. Picking  $K$  too small is a much bigger mistake than picking  $K$  too large. You cannot get three modes out of a mixture of two normals, but you can essentially ignore extra mixture terms and get bimodal distributions from mixtures of three or more distributions.

Using independent reference priors  $\tau_i \sim \text{Gamma}(0.001, 0.001)$ ,  $\mu_i \sim N(0, 1000)$ , and  $(w_1, \dots, w_K) \sim \text{Dir}(1/K, \dots, 1/K)$ , the LPMLs for  $K = 3, 4, 5, 6, 7$  are  $-193.2$ ,  $-184.0$ ,  $-180.4$ ,  $-175.5$ , and  $-178.5$ , indicating  $K = 6$  provides the best predictive fit. Figure 15.3 illustrates the results of fitting some mixture models. A simple normal model ( $K = 1$ ) gives an LPML of  $-243.1$  and was illustrated in Figure 15.1.

The number of mixands  $K$  can be chosen *a priori* to be quite large. The number of actual components being used can be monitored through an MCMC run and the analysis modified accordingly. The selection of  $K$  to a number a bit larger than is necessary, but not grossly more, achieves parsimony and can improve the predictive ability of the model.

When the densities  $\phi_k$  are the same, it is often reasonable to think of the vectors  $\theta_k$  as exchangeable in that their ordering should not matter. Depending on the situation, the  $w_k$ s may also be exchangeable. A general, yet convenient, prior for a mixture of normals is

$$\mu_1, \mu_2, \dots, \mu_K \stackrel{iid}{\sim} N(\tilde{m}, 1/\tilde{\tau}) \quad \perp \!\!\! \perp \quad \tau_1, \tau_2, \dots, \tau_K \stackrel{iid}{\sim} \text{Gamma}(a, b),$$

with independent weights

$$(w_1, w_2, w_3, \dots, w_K) \sim \text{Dir}(\alpha/K, \dots, \alpha/K).$$

To be useful, this prior needs some additional restrictions as discussed later. If desired, hyperpriors can be placed on  $\tilde{m}$ ,  $\tilde{\tau}$ ,  $a$ ,  $b$ , and  $\alpha$ .

Without further qualification or restriction, mixture models often lack identifiability. This means that there exists more than one set of parameter values that will generate the same distribution for the data, cf. Section 4.14 and the Hui-Walter model of Chapter 14. With an exchangeable prior like

that just presented, lack of identifiability can result in serious problems that we discuss near the end of this subsection. Fortunately, the identifiability problem is often resolvable. In a mixture of two normals, if we constrain the means so that  $\mu_1 < \mu_2$ , the model becomes identifiable. In the  $K$  normals mixture, placing an ordering on all  $K$  means similarly resolves the issue. Ordering the means is sufficient; we need impose no restrictions on the precisions. For a Bayesian, the simplest way to impose an ordering on the means is to specify a prior that is ordered with probability 1. Imposing such an ordering is easy in WinBUGS, see the code in Example 15.1.3. But just restricting the model to be identifiable does not guarantee well behaved Markov chains!

Nonidentifiability is mitigated if an informative prior is incorporated into the analysis. For example, if the data are people's heights, we can construct informative priors for both males and females and typically the prior for male heights would be (stochastically) larger. (We might be sampling from a mixture of female basketball players and male gymnasts.) Thus, our information on females provides an informative prior on  $(\mu_1, \tau_1)$  and our information on males provides an informative prior on  $(\mu_2, \tau_2)$ . The informative prior should reduce identifiability problems, but there seems little reason not to require  $\mu_2 > \mu_1$ , which definitely eliminates any identifiability problems. We might also have prior information on the proportion of females in the sample. Mixture distributions can be difficult to fit and, when it is available, good prior information certainly helps.

To analyze a random sample of mixture data, say,  $y_1, \dots, y_n$ , we introduce independent subpopulation variables  $X_{ik}$  with  $X_{ik} \sim \phi_k(y|\theta_k)$  for  $k = 1, \dots, K$  and independent index variables  $W_i$  with  $\Pr[W_i = k] = w_k$ . The random observation from the mixture is now

$$y_i = X_{iW_i} = \sum_{k=1}^K X_{ik} I_{\{k\}}(W_i).$$

The first equality provides a particularly convenient way to code mixtures in WinBUGS using the `dcat` option. Note that for  $K = 2$ ,  $(W_i - 1) \sim \text{Bern}(w_2)$ .

**EXAMPLE 15.1.3.** The data  $y$  are a simulated random sample from a mixture of two distributions with  $n = 100$ , specifically  $N(0, 1)$  and  $N(3, 1/2)$  distributions with  $w_1 = 0.34$ . The complete data are on our website. Here  $W_i$  takes the values  $(1, 2)$  with probabilities  $\{(w_1, w_2) : w_2 = 1 - w_1\}$ ;  $W_i = 1$  indicates that  $y_i$  was taken from the first population, and  $W_i = 2$  indicates it was taken from the second. Our WinBUGS code is

```
model{
  for(i in 1:n){
    W[i] ~ dcat(w[1:2])
    y[i] ~ dnorm(mu[W[i]], tau[W[i]])
  }
  w[1:2] ~ ddirch(alpha[1:2])
  mu[1] ~ dnorm(0, 0.0001)
  mu[2] ~ dnorm(0, 0.0001)I(mu[1],)
  for(i in 1:2){ tau[i] ~ dgamma(0.001, 0.001) }
}
list(w=c(0.3, 0.7), mu=c(0, 2))
list(n=100, alpha = c(1, 1),
  y = c(-1.902, -1.579, -1.368, -1.204,
        -1.068, -0.949, -0.842, -0.744,
        -0.652, -0.566, ..., 4.883, 5.172))
```

The prior has  $\mu_1 \sim N(0, 10000)$  and, using a WinBUGS feature discussed in Chapter 12,  $\mu_2 | \mu_1 \sim N(0, 10000)I_{(\mu_1, \infty)}(\mu_2)$ , so that  $\mu_2$  is  $N(0, 10000)$  but conditioned on  $\mu_2 > \mu_1$ . Our MCMC initial value for  $\mu_2$  is larger than for  $\mu_1$ . Unfortunately, we got a *trap* message using related code. If you get trapped, close the *trap* window and click on the *update* button. You may be able to get a

reasonable sample this way. Alternatively, you could change the initial value for  $\mu_2$ , increase its prior mean, or use a random walk prior as discussed after Exercise 15.2.

**EXERCISE 15.2.** (a) Run the code from Example 15.1.3 to obtain estimates of the means, precisions, and the mixing parameter. (b) Modify the code to obtain the predictive density of a future observation from the model (see the code in Exercise 15.3). Take the WinBUGS output from CODA and make a density plot in R. Plot the true mixture density on the same graph and evaluate the estimate. (c) Now run the code with different priors including one with  $\mu_1 \sim N(0, 1)$ ,  $\mu_2 | \mu_1 \sim N(2, 1)I_{(\mu_1, \infty)}(\mu_2)$  and independently  $\tau_k \sim \text{Gamma}(1, 1)$ . Report on what priors you tried, what worked, and what didn't. In no way are we advocating trial and error for obtaining priors. We are illustrating the importance of good prior information for mixture problems.

An alternative to the exchangeable prior with mean restrictions is to use a *random walk prior*. In this prior  $\mu_{k+1} = \mu_k + \delta_k$  where  $\delta_k$  a positive random variable. A reference prior for the mixture model places a reference prior on  $\mu_1$  and independent reference priors on each  $\delta_k$ . In Exercise 15.3,  $K = 2$  and we take  $\delta_1$  as a half-normal distribution. Specifically, our prior has  $\mu_1 \sim N(0, 1000)$  and  $\mu_2 = \mu_1 + \delta_1$  where  $\delta_1 \sim N(0, 1000)$  constrained so that  $\delta_1 > 0$ ; hence  $\mu_2 > \mu_1$ . The exercise also involves real data and illustrates computation of the predictive density.

**EXERCISE 15.3.** Consider  $n = 66$  log-transformed ELISA scores from cows infected with Johnes' disease (Choi et al., 2006a). The authors analyzed these data using a simple normal model. Assume instead that the data  $y_i$  are a random sample from a mixture of two normal distributions. In WinBUGS use

```
model{
  # Sampling model
  for(i in 1:n){
    W[i] ~ dcat(w[1:2])
    y[i] ~ dnorm(mu[W[i]], tau[W[i]])
  }
  # Prior
  w[1:2] ~ ddirch(alpha[1:2])
  mu[1] ~ dnorm(0, 0.001)
  delta1 ~ dnorm(0, 0.001)I(0, )
  mu[2] <- mu[1]+delta1
  for(i in 1:2){tau[i] ~ dgamma(0.001, 0.001)}
  # Density estimate
  for(i in 1:100){
    grid[i] <- -3.5+5*(i-1)/99
    f[i] <- 0.3989*w[1]*sqrt(tau[1])
      *exp(-0.5*pow(grid[i]-mu[1], 2)*tau[1])
      +0.3989*w[2]*sqrt(tau[2])
      *exp(-0.5*pow(grid[i]-mu[2], 2)*tau[2])
  }
}
list(w=c(0.5, 0.5), mu=c(-2, NA), delta1=2, tau=c(1, 1)) # inits
list(n=66, alpha=c(1, 1), y=c(-2.3, 0.66, -1.11, 0.52, 0.93, 0.8, 0.07,
  0.43, -2.66, 0.46, -0.34, 0.53, 0.17, -0.04, 0.35, -2.81, 0.4, 0.13, 0.88,
  0.83, -2.3, -2.53, -1.35, -1.71, 0.55, -3.22, 0.13, 0.6, -0.07, -1.56, 1.02,
  0.61, 0.52, 0.03, 0.6, -2.04, 0.76, -1.39, 0.51, 0.15, -0.26, 0.98, 0.8,
  0.24, 0.12, 0.67, -0.07, -0.31, 0.41, -1.05, -1.35, -1.71, 0.92, -1.2, -2.3,
  -1.27, 0.49, -0.26, -2.81, -0.82, -0.73, -0.2, 0.59, 0.58, -1.77, -2.04))
```

(a) Run the code and obtain inferences for the means, precisions, and the mixing parameter. Plot the fitted density on top of a histogram of the raw data. (b) Repeat this exercise with a simple (one-component) normal model; comment on the fits of the one-component versus two-component models.

#### 15.1.1.1 Identifiability Issues\*

Mixture models are notorious for presenting identifiability problems. The general concept of identifiability was discussed in Section 4.14 including an example of a mixture of two normals with known variances. This discussion presupposes familiarity with the earlier one.

Imagine sampling from a mixture of  $K$  subpopulations, each characterized with a density  $\phi$  involving an identifiable vector parameter  $\theta_k$  of dimension  $r$  and let  $w_k$  be the proportion of the overall population that comes from the  $k$ th subpopulation. Let  $X_k$  denote an observation from the  $k$ th subpopulation and let  $W$  randomly determine the subpopulation sampled. Specifically, for  $k = 1, \dots, K$ , consider the well-defined (identifiable) sampling model for all subpopulations

$$X_k | \theta_k \stackrel{\text{ind}}{\sim} \phi(x|\theta_k) \quad \perp\!\!\!\perp W | w_1, \dots, w_K \sim \Pr[W=k] = w_k.$$

The  $w_k$ s must be nonnegative and sum to 1. For the complete set of random variables  $Z \equiv (X_1, \dots, X_K, W)'$ , write the parameter vector as  $\Theta \equiv (\theta'_1, \dots, \theta'_K, w_1, \dots, w_K)'$ .  $\Theta$  is identifiable for  $Z$ . The  $K$  normals mixture model is the special case that takes  $\theta_k = (\mu_k, \tau_k)'$  and each  $\phi(y|\mu_k, \tau_k)$  to be the density of a  $N(\mu_k, 1/\tau_k)$ .

Unfortunately, in the mixture model an observation is randomly chosen from one subpopulation, so we only observe

$$y = \sum_{k=1}^K X_k I_{\{k\}}(W).$$

By considering how to compute  $\Pr[y \in A]$  for some set  $A$ , it is easy to see that the density of  $y$  is

$$f(y|\Theta) \equiv f(y|\theta_1, \dots, \theta_K; w_1, \dots, w_K) = \sum_{k=1}^K w_k \phi(y|\theta_k).$$

Just as for the mixture of two normals in Section 4.14, the originally identifiable parameter  $\Theta$  for the vector  $Z$  becomes nonidentifiable when only  $y$  is observed.

For the  $K$  normals mixture model, one often assumes  $\mu_1 < \dots < \mu_K$ . Specifically,  $\Theta = [(\mu_1, \tau_1, w_1), \dots, (\mu_K, \tau_K, w_K)]$  is the original parameter vector. To obtain an identifiable parameterization for  $y$ , consider  $y|g(\Theta)$  where  $g(\Theta) = [(\mu_{(1)}, \tau^{(1)}, w^{(1)}), \dots, (\mu_{(K)}, \tau^{(K)}, w^{(K)})]$  with  $\mu_{(1)} < \mu_{(2)} < \dots < \mu_{(K)}$  and if  $\mu_{(k)} = \mu_j$ , then  $\tau^{(k)} \equiv \tau_j$  and  $w^{(k)} \equiv w_j$ . Another common, and quite general, way to impose identifiability in mixture models is to require  $w_1 < w_2 < \dots < w_K$ .

Restricting the parameters of the sampling distribution makes the corresponding likelihood painful to use. It is difficult to maximize for frequentists and it is difficult to integrate for Bayesians. However, using MCMC, Bayesians have a convenient way of avoiding this pain. After finding an appropriate restriction on the parameters that is identifiable, it can be relatively easy to sample from a prior distribution that gives probability 1 to parameter values that satisfy the restriction. So rather than actually restricting the parameters of the sampling distribution, it is often easier to let the prior do the work by defining a prior that satisfies those same restrictions with probability 1. In a  $K$  normals mixture model, rather than assuming  $\mu_1 < \dots < \mu_K$  in the sampling distribution, we can choose a prior with  $\Pr[\mu_1 < \dots < \mu_K] = 1$  to accomplish the same thing. In general, this would often occur by defining a prior on  $\Theta$  but restricting it to  $g(\Theta)$ . For example, the exchangeable prior for normal mixtures is restricted to ordered means.

As discussed in Section 4.14, it is tempting for Bayesians to be careless about identifiability because answers can be obtained without it. In particular, informative priors often eliminate or mitigate identifiability problems. For example, an informative prior that happens to have a large

probability for  $\Pr[\mu_1 < \dots < \mu_K]$  will mitigate identifiability problems in  $K$  normals mixtures even if the user is ignoring them.

Priors can help with identifiability problems, but they can also exacerbate them. One might assume a prior with

$$\theta_1, \dots, \theta_K \perp\!\!\!\perp w_1, \dots, w_K$$

and, as discussed above, it often seems reasonable to think of the vectors  $\theta_k$  as exchangeable in that their ordering should not matter, so assume

$$\theta_1, \theta_2, \dots, \theta_K \stackrel{iid}{\sim} p(\theta).$$

Then, given our assumptions, for any  $k$

$$y|W=k \sim X_k \sim \int \phi(y|\theta) p(\theta) d\theta$$

which is a distribution that does not depend on  $k$ . Thus the data  $y$  are independent of  $W$  and there is no information to be gained about  $W$  from the data. In the normal case with  $X_k|\theta_k \sim N(\theta_k, 1)$  and  $\theta_k \sim N(0, 1)$ , we get  $X_k \sim N(0, 2)$  for all  $k$ . Since  $y$  comes from one of these  $K$  indistinguishable distributions, it is no wonder  $y$  cannot give information on which subpopulations are more likely to have generated it.

Marin and Robert (2007) contains a good discussion of Bayesian finite mixture models.

### 15.1.2 Dirichlet Process Mixtures: Infinite Mixtures

In the previous subsection, we considered finite mixtures

$$f(y|\theta_1, \dots, \theta_K; w_1, \dots, w_K) = \sum_{k=1}^K w_k \phi(y|\theta_k).$$

We now look at infinite mixtures

$$f(y|\theta_1, \theta_2, \dots; w_1, w_2, \dots) = \sum_{k=1}^{\infty} w_k \phi(y|\theta_k). \quad (2)$$

One way to generate a density such as (2) with fixed values for  $w_k$  and  $\theta_k$  is to think of the sampling distribution  $\phi(y|\theta)$  and place a discrete prior distribution  $G$  on  $\theta$  with density  $p_G(\theta_k) = w_k$ ,  $k = 1, 2, \dots$ . The marginal distribution of  $y$  is then,

$$f(y) = \int \phi(y|\theta) dG(\theta) \equiv \sum_{k=1}^{\infty} w_k \phi(y|\theta_k) \quad (3)$$

where the notation  $dG(\theta)$  turns the integral into the appropriate sum because the distribution is discrete.  $G$  depends on the values chosen for  $\theta_1, \theta_2, \dots$  and  $w_1, w_2, \dots$ . We can think of these as hyperparameters of the prior distribution  $G$  and rewrite (3) as

$$f(y|\theta_1, \theta_2, \dots; w_1, w_2, \dots) \equiv f(y|G) = \int \phi(y|\theta) dG(\theta) = \sum_{k=1}^{\infty} w_k \phi(y|\theta_k).$$

The next thing we are going to do is put a prior on these (hyper)parameters, thus creating a hierarchical model as in Section 4.12.

Placing priors on these parameters amounts to picking a discrete distribution  $G$  that is chosen at random. Picking a random distribution sounds strange, but we do it for the  $y$ s all the time. If  $y|\theta$  has density  $f(y|\theta)$  and  $\theta$  has a prior distribution with density  $p(\theta)$ , we are conceptually picking a  $\theta$  at

random and applying the random sampling distribution with density  $f(y|\theta)$ , so we have a random distribution for  $y$ . Our usual marginal distribution is an average of these random distributions,

$$f(y) = \int f(y|\theta)p(\theta)d\theta,$$

which is a form of mixture distribution. Similarly, when we have data available, our usual predictive distribution for  $\tilde{y}$  given data  $y$  is also a mixture,

$$f_p(\tilde{y}|y) = \int f_p(\tilde{y}|\theta)p(\theta|y)d\theta,$$

again obtained by averaging over a random density  $f_p(\tilde{y}|\theta)$  with a distribution on  $\theta$ .

We are just moving the procedure back a step. Now we are picking a random distribution  $G$  for  $\theta$  by putting a prior on  $\theta_1, \theta_2, \dots; w_1, w_2, \dots$ . Placing a reasonable prior on the  $\theta$ s is easy. Simply take them as iid from a single distribution that is our best prior guess for the distribution of  $\theta$ , call this  $G_0$ . However, putting a prior on the  $w$ s is more difficult because they have to be randomly chosen numbers that are between 0 and 1 that add to 1.

Rather than putting a prior on the sequence of probabilities  $w_1, w_2, w_3, \dots$ , we instead put a prior on another sequence  $q_1, q_2, q_3, \dots$  that just consists of numbers between 0 and 1. From this we induce a prior on  $w_1, w_2, w_3, \dots$  by the transformation

$$w_k = q_k \prod_{j=1}^{k-1} (1 - q_j) \quad (4)$$

with inverse transformation

$$q_k = w_k / \left( 1 - \sum_{j=1}^{k-1} w_j \right). \quad (5)$$

For (5) to be the inverse of (4) requires  $(1 - \sum_{j=1}^{k-1} w_j) = \prod_{j=1}^{k-1} (1 - q_j)$ . This is easily proved by induction after observing from (5) that  $1 - q_k = (1 - \sum_{j=1}^k w_j) / (1 - \sum_{j=1}^{k-1} w_j)$ . The choice of the transformation (4) is discussed below. In particular, we take

$$q_i \stackrel{iid}{\sim} \text{Beta}(1, \alpha).$$

The question now becomes, “How do we know that these random  $w$ s define a set of probabilities, i.e., are nonnegative and sum to 1?” A proof is given on our website.

This approach to defining a prior distribution on a sequence of weights  $w_1, w_2, \dots$  can be thought of as breaking a stick of length 1. Each  $w_k$  is the length of a piece we break off. The  $w$ s should add up to 1. First we break off a fraction  $q_1 \sim \text{Beta}(1, \alpha)$  from the stick. The length of the broken off piece is  $w_1 = q_1$ . There is  $1 - w_1$  left of the stick. Now break off the fraction  $q_2 \sim \text{Beta}(1, \alpha)$  of what remains. The length of this new piece is  $w_2 = q_2(1 - w_1) = q_2(1 - q_1)$ . Again, break off a fraction  $q_3 \sim \text{Beta}(1, \alpha)$ . The length is  $w_3 = q_3(1 - w_1 - w_2)$ . This is essentially equation (5) with  $k = 3$ , so by (4)  $w_3 = q_3(1 - q_1)(1 - q_2)$ . Continuing this process gives

$$w_k = q_k \prod_{j=1}^{k-1} (1 - q_j), \quad q_1, q_2, \dots \stackrel{iid}{\sim} \text{Beta}(1, \alpha).$$

Taking

$$\theta_1, \theta_2, \dots \stackrel{iid}{\sim} G_0$$

together with the distribution of the  $w$ s, these determine our random distribution  $G$  by placing a prior on all of  $G$ ’s hyperparameters. The prior on  $G$  depends on only two parameters: the weight  $\alpha$  in the Beta distributions and the distribution  $G_0$ .

This method for determining a random probability distribution  $G$  is known as a *Dirichlet process* (*DP*) and is written

$$G \sim \mathcal{D}(\alpha, G_0).$$

It was introduced by Ferguson (1973, 1974). The stick-breaking representation is due to Sethuraman (1994). Dirichlet process methods have dominated much of modern Bayesian nonparametric modeling.

Dirichlet processes have two particularly nice properties. First, if you take one observation  $\tilde{\theta}$  with cdf  $G(\theta|\theta_1, \theta_2, \theta_3, \dots, w_1, w_2, w_3, \dots)$  and the  $\theta_k$ s and  $w_k$ s are randomly chosen as described here so that  $G \sim \mathcal{D}(\alpha, G_0)$ , then the marginal distribution of  $\tilde{\theta}$  is just the  $G_0$  distribution. Symbolically, if

$$\tilde{\theta}|G \sim G; \quad G \sim \mathcal{D}(\alpha, G_0)$$

then

$$\tilde{\theta} \sim G_0.$$

Second, Dirichlet processes are closed under sampling. In other words, the posterior distribution is also a Dirichlet process. Specifically, suppose that  $\tilde{\theta}_1, \dots, \tilde{\theta}_n$  is a random sample with the cdf  $G(\theta|\theta_1, \theta_2, \theta_3, \dots, w_1, w_2, w_3, \dots)$  and with  $\theta_k$ s and  $w_k$ s chosen so that  $G \sim \mathcal{D}(\alpha, G_0)$ , then we write

$$\tilde{\theta}_1, \dots, \tilde{\theta}_n|G \stackrel{iid}{\sim} G; \quad G \sim \mathcal{D}(\alpha, G_0).$$

It turns out that

$$G|\tilde{\theta}_1, \dots, \tilde{\theta}_n \sim \mathcal{D}\left(\alpha + n, \frac{\alpha}{\alpha + n}G_0 + \frac{1}{\alpha + n} \sum_{j=1}^n \delta_{\tilde{\theta}_j}\right)$$

where  $\delta_{\tilde{\theta}_j}$  is the distribution that gives probability 1 to  $\tilde{\theta}_j$ . Clearly small values of  $\alpha$  let the data play a larger role.

Sampling from (2) with the mixture determined by a Dirichlet process is called sampling from a *Dirichlet process mixture* (*DPM*) distribution. This has its roots in Antoniak (1974), Lo (1984), and others. A DPM density is written hierarchically as

$$f(y|G, \phi) \equiv \int \phi(y|\theta) dG(\theta) \equiv \sum_{k=1}^{\infty} w_k \phi(y|\theta_k),$$

$$G \sim \mathcal{D}(\alpha, G_0).$$

Dirichlet process mixtures involve an infinite mixture of  $\phi(y|\theta)$ s but we can only compute finite mixtures. Simple approximations truncate the sequences  $\theta_1, \theta_2, \theta_3, \dots$  and  $w_1, w_2, w_3, \dots$  for some large value of  $K$  into  $\theta_1, \theta_2, \theta_3, \dots, \theta_K$  and  $w_1, w_2, w_3, \dots, w_K$ . Thus finite approximations to the Dirichlet process use a distribution  $G_K$  that has  $\Pr[\theta = \theta_k] = w_k$  for  $k = 1, \dots, K$  and define distributions for the  $\theta_k$ s and  $w_k$ s. A DPM is approximated similar to (3),

$$f(y|G_K) = f(y|\theta_1, \dots, \theta_K, w_1, \dots, w_K) = \int \phi(y|\theta) dG_K(\theta) = \sum_{k=1}^K w_k \phi(y|\theta_k).$$

To make  $G_K$  random, it is natural to take  $\theta_k \stackrel{iid}{\sim} G_0$ . Another natural approximation defines the distribution of  $w_1, w_2, w_3, \dots, w_{K-1}$  as in a DP but then takes  $w_K \equiv 1 - \sum_{k=1}^{K-1} w_k$ . Ishwaran and Zarepour (2002) [IZ] discuss a different approximation to the DP that has been used by many authors and that is simpler to compute. They take  $(w_1, \dots, w_K) \sim \text{Dirich}(\alpha/K, \dots, \alpha/K)$ . IZ argue that their distribution for  $G_K$  converges to the Dirichlet process  $\mathcal{D}(\alpha, G_0)$  as  $K \rightarrow \infty$ . This is by no means obvious. We present an heuristic justification.

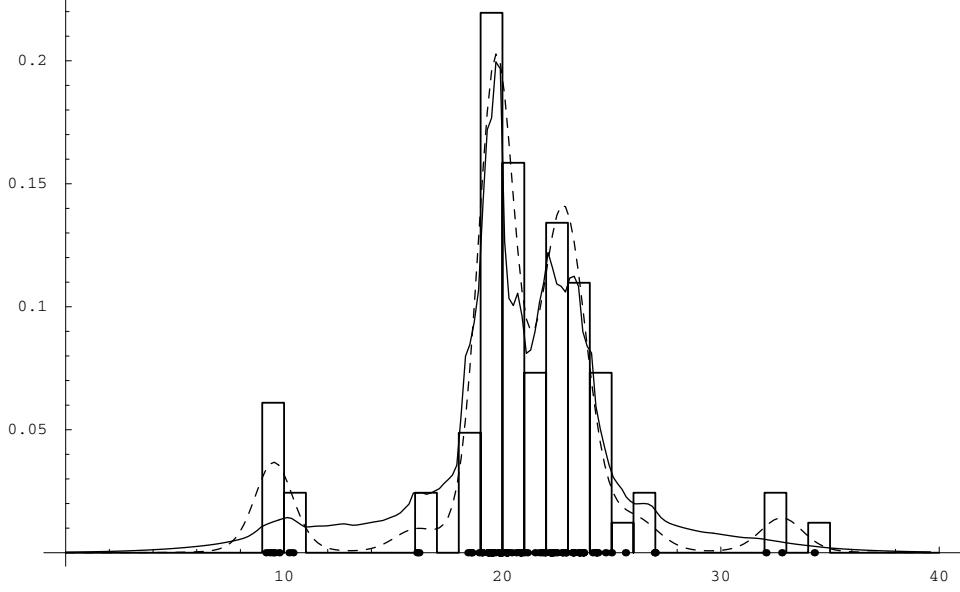


Figure 15.4: Galaxy data: Dirichlet process mixture (dashed) and mixture of Polya trees (solid) fits.

The IZ approximation is based on the original Ferguson (1973) definition of the Dirichlet process. By definition,  $G \sim \mathcal{D}(\alpha, G_0)$  if for any  $J$  and any partition of the sample space  $\{A_j : j = 1, \dots, J\}$ , the random probabilities of the partition sets satisfy

$$[G(A_1), G(A_2), \dots, G(A_J)] \sim \text{Dirich}[\alpha G_0(A_1), \alpha G_0(A_2), \dots, \alpha G_0(A_J)].$$

Keeping the  $\theta_k$ s fixed in  $G_K$ , place a  $\text{Dirich}(\alpha/K, \dots, \alpha/K)$  distribution on the  $w_k$ s to make  $G_K$  random. With the *empirical distribution*  $F_K = (1/K) \sum_{k=1}^K \delta_{\theta_k}$ , it is not difficult to see that for any partition

$$\begin{aligned} [G_K(A_1), G_K(A_2), \dots, G_K(A_J)] &\Big| \theta_1, \theta_2, \theta_3, \dots, \theta_K \\ &\sim \text{Dirich}[\alpha F_K(A_1), \alpha F_K(A_2), \dots, \alpha F_K(A_J)], \end{aligned}$$

so by Ferguson's definition

$$G_K | \theta_1, \theta_2, \theta_3, \dots, \theta_K \sim \mathcal{D}(\alpha, F_K).$$

Finally, if the  $\theta_k$ s are iid from  $G_0$ , by standard results from probability theory, the empirical distribution  $F_K$  tends to  $G_0$  as  $K$  grows. So heuristically,  $G_K$  is approximately distributed as  $\mathcal{D}(\alpha, G_0)$  for large  $K$ .

**EXAMPLE 15.1.4. Galaxy Data.** Consider a Dirichlet process mixture of normals with  $\alpha = 1$  and  $G_0$  taken as a bivariate distribution with independent components. The first variable corresponds to the mean and the second to the precision of the normal distributions. We used a reference prior with  $N(0, 1000)$  for the mean and a  $\text{Gamma}(0.001, 0.001)$  for the precision. The IZ Dirichlet process approximation with  $K = 50$  takes independent samples  $\mu_k \sim N(0, 1000)$ ,  $\tau_k \sim \text{Gamma}(0.001, 0.001)$ , and  $(w_1, \dots, w_K) \sim \text{Dir}(1/50, \dots, 1/50)$ . Note the similarity to Example 15.1.2. For fitting this model, LPML is  $-158.4$  which is considerably better than the LPML values for the low dimensional mixtures. The fit is illustrated in Figure 15.4 along with the mixture of Polya trees fitted model developed in the next subsection.

**EXERCISE 15.4.** WinBUGS code for fitting Example 15.1.4 is available in `DPMdensity.odc` on the book website. The code includes lines for computing the CPO statistics as well as the predictive density over a grid of points. Use the code to duplicate the Dirichlet process mixture of normals fit in Figure 15.4. Rerun the code with  $\alpha = 0.1$ ,  $\alpha = 1$ , and  $\alpha = 10$ , being sure to monitor the number of “active” components via the node `total`. Do the predictive density, LPML, and number of components change very much with  $\alpha$ ?

DPpackage provides a slightly different analysis for Example 15.1.4.

**EXERCISE 15.5.** Install DPpackage into your version of R, cf. Appendix C. The function `DPdensity` fits a Dirichlet process mixture of normal distributions as proposed by Escobar and West (1995). Type `help(DPdensity)` in R to see a description of the model. How does this model differ from Example 15.1.4? Use this DP function to obtain a plot similar to Figure 15.4. Code with a “default” prior specification is found in `Chap15DPpackage.txt` on our website. The `DPdensity` function allows for random  $\alpha$ . Try  $\alpha \sim \text{Gamma}(1, 1)$  and  $\alpha \sim \text{Gamma}(1, 0.1)$ . How do posterior inferences change, especially the number of clusters (`ncluster`), the predictive density, and the LPML?

### 15.1.3 Mixtures of Polya Trees

Our discussion follows Christensen, Hanson, and Jara (2008). The general definition of mixtures of Polya trees (MPTs), such as that in Lavine (1992, 1994), is quite broad. Using the simpler definition in Hanson (2006), we use Polya trees to generalize the  $N(\mu, \sigma^2)$  family of distributions, see Figure a. Other parametric families are generalized similarly.

The generalization goes through a number of stages, say  $J$ . At each stage we introduce new parameters to generalize the previous stage. At the first stage, we split the real number line, that is, the support of the normal distribution, into two intervals divided by the median  $\mu$ . We then allow changes in the probabilities of being below or above  $\mu$  but we retain the shape of the normal density both below  $\mu$  and above  $\mu$ . Figure b illustrates the density for the case when the probability of being below  $\mu$  is 0.45.

The new parameters at the first stage are  $\theta_{11}$ , the probability of being no greater than  $\mu$ , and  $\theta_{12}$ , the probability of being above  $\mu$ . Formally, let  $X_1$  have the first stage distribution, then

$$\theta_{11} \equiv \Pr[X_1 \leq \mu]$$

and

$$\theta_{12} \equiv \Pr[X_1 > \mu] = 1 - \theta_{11}.$$

Because we retain the shape of the normal on both sets, if  $a \leq \mu$  and  $Y \sim N(\mu, \sigma^2)$ , conditionally we have

$$\Pr[X_1 \leq a | X_1 \leq \mu] \equiv \frac{\Pr[Y \leq a]}{0.5} = 2\Phi[(a - \mu)/\sigma]$$

where  $\Phi(\cdot)$  is the cdf of a standard normal. Similarly, if  $b > \mu$ ,

$$\Pr[X_1 > b | X_1 > \mu] \equiv 2\Pr[Y > b] = 2\{1 - \Phi[(b - \mu)/\sigma]\}.$$

Alternatively, we can write

$$\begin{aligned}\Pr[X_1 \leq a] &= \Pr[X_1 \leq a | X_1 \leq \mu] \Pr[X_1 \leq \mu] = \Pr[Y \leq a] 2\theta_{11} \\ \Pr[X_1 > b] &= \Pr[X_1 > b | X_1 > \mu] \Pr[X_1 > \mu] = \Pr[Y > b] 2\theta_{12}.\end{aligned}$$

The density of the stage 1 distribution is

$$f(x_1 | \mu, \sigma^2, \theta_{11}, \theta_{12}) = \frac{2^1}{\sqrt{2\pi}\sigma} e^{-(x_1 - \mu)^2 / 2\sigma^2} [\theta_{11} I_{(-\infty, \mu]}(x_1) + \theta_{12} I_{(\mu, \infty)}(x_1)].$$

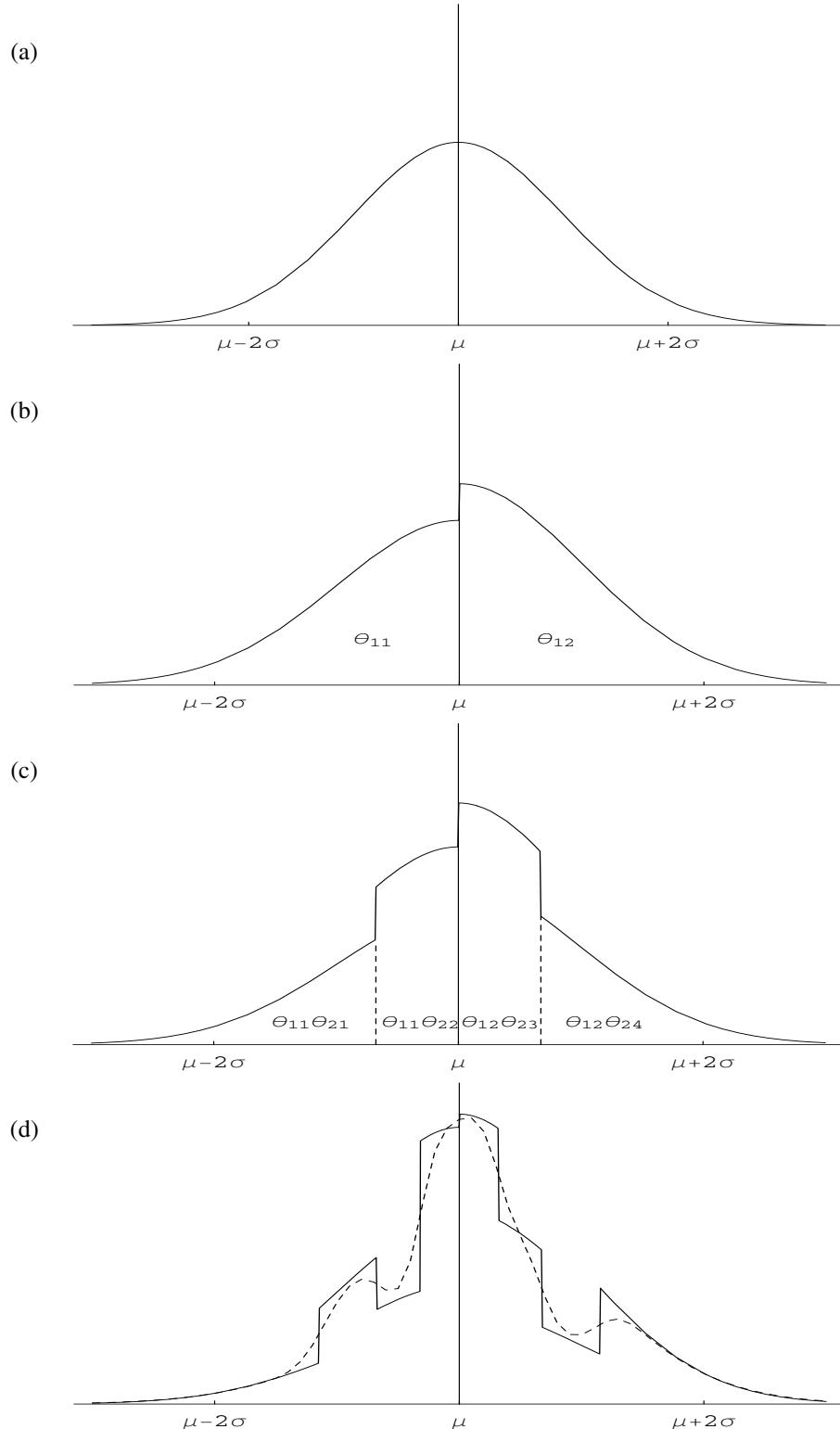


Figure 15.5: *Finite PT densities.* (a) First:  $N(\mu, \sigma^2)$  centering density. (b) Second:  $\theta_{11} = 0.45$ . (c) Third:  $\theta_{21} = 0.4$ ,  $\theta_{23} = 0.6$ . (d) Fourth:  $\theta_{31} = 0.3$ ,  $\theta_{33} = 0.3$ ,  $\theta_{35} = 0.6$ ,  $\theta_{37} = 0.3$ ; centering family mixed over  $\mu \sim N(\mu_0, (\sigma/10)^2)$ ,  $\sigma \sim N(\sigma_0, (\sigma_0/10)^2)$ .

In the first stage of the process, we split the real line at the median  $\mu$  of the  $N(\mu, \sigma^2)$  distribution. For the second stage, we split the line at the quartiles, say,  $q_1, \mu, q_3$  of the original distribution leading us to consider the sets  $(-\infty, q_1], (q_1, \mu], (\mu, q_3], (q_3, \infty)$ . For the normal, the quartiles are  $q_1 \equiv \mu - 0.6745\sigma$ ,  $\mu$ , and  $q_3 \equiv \mu + 0.6745\sigma$ . Under the original distribution, each of these sets has probability 0.25, but in stage 2 we allow the probabilities of the sets to change in a manner that is consistent with stage 1. An illustration of a stage 2 density is Figure c. Letting  $X_2$  have the second stage distribution, we introduce new parameters,  $\theta_{21}, \theta_{22}, \theta_{23}, \theta_{24}$ , defined as conditional probabilities relative to the sets used in stage 1:

$$\begin{aligned}\theta_{21} &= \Pr[X_2 \leq q_1 | X_2 \leq \mu] \\ \theta_{22} &= \Pr[q_1 < X_2 \leq \mu | X_2 \leq \mu] \\ \theta_{23} &= \Pr[\mu < X_2 \leq q_3 | X_2 > \mu] \\ \theta_{24} &= \Pr[q_3 < X_2 | X_2 > \mu].\end{aligned}$$

Note that  $\theta_{21} = 1 - \theta_{22}$  and  $\theta_{23} = 1 - \theta_{24}$ . Unconditionally, the four sets have the probabilities

$$\begin{aligned}\Pr[X_2 \leq q_1] &= \theta_{11}\theta_{21} \\ \Pr[q_1 < X_2 \leq \mu] &= \theta_{11}\theta_{22} \\ \Pr[\mu < X_2 \leq q_3] &= \theta_{12}\theta_{23} \\ \Pr[q_3 < X_2] &= \theta_{12}\theta_{24}.\end{aligned}$$

Within each set, we again use the shape of the original normal density so, for example, if  $\mu < a < b \leq q_3$  and  $Y \sim N(\mu, \sigma^2)$ ,

$$\begin{aligned}\Pr[a < X_2 \leq b] &= \Pr[a < X_2 \leq b | \mu < X_2 \leq q_3]\Pr[\mu < X_2 \leq q_3] \\ &= \Pr[a < Y \leq b | \mu < Y \leq q_3]\Pr[\mu < X_2 \leq q_3] \\ &= \frac{\Pr[a < Y \leq b]}{\Pr[\mu < Y \leq q_3]}\Pr[\mu < X_2 \leq q_3] \\ &= \Pr[a < Y \leq b] \frac{\theta_{12}\theta_{23}}{0.25}.\end{aligned}$$

In Figure c the density has  $\theta_{11} = 0.45$ ,  $\theta_{21} = 0.4$ ,  $\theta_{23} = 0.6$ . In general, the density of the stage 2 distribution is

$$f(x_2 | \mu, \sigma^2, \theta_{11}, \theta_{12}, \theta_{21}, \theta_{22}, \theta_{23}, \theta_{24}) = \frac{2^2}{\sqrt{2\pi}\sigma} e^{-(x_2 - \mu)^2 / 2\sigma^2} \times \\ [\theta_{11}\theta_{21}I_{(-\infty, q_1]}(x_2) + \theta_{11}\theta_{22}I_{(q_1, \mu]}(x_2) + \theta_{12}\theta_{23}I_{(\mu, q_3]}(x_2) + \theta_{12}\theta_{24}I_{(q_3, \infty)}(x_2)].$$

Subsequent stages follow a similar pattern with stage 3 breaking the support of the  $N(\mu, \sigma^2)$  distribution into eight sets based on the octiles so that each set has probability  $1/8 = 1/2^3$  under the original parametric distribution. One introduces parameters  $\theta_{31}, \dots, \theta_{38}$  for the conditional probabilities of these sets given the stage 2 sets. Each parameter whose second subscript is even equals 1 minus the previous parameter, for example  $\theta_{32} = 1 - \theta_{31}$  and  $\theta_{38} = 1 - \theta_{37}$ . Figure d illustrates stage 3 for  $\theta_{31} = 0.3$ ,  $\theta_{33} = 0.3$ ,  $\theta_{35} = 0.6$ ,  $\theta_{37} = 0.3$ , and the previous  $\theta_{js}$ s. One continues these stages to a level  $J$  with  $2^J$  sets that each have probability  $1/2^J$  under the original parametric distribution and whose conditional probabilities given the stage  $J - 1$  sets are the parameters  $\theta_{J1}, \dots, \theta_{J,2^J}$  in which  $\theta_{J,2k-1} = 1 - \theta_{J,2k}$ ,  $k = 1, \dots, 2^{J-1}$ .

At the final stage  $J$ , the density at a point  $x_J$  depends on the string of sets from the various stages that contain  $x_J$ . For example, with  $J = 3$ , if  $x_3$  is between the fifth and sixth octiles, that is, if  $\mu + 0.3186\sigma < x_3 \leq \mu + 0.6745\sigma$ , then  $x_3$  is also in the sets  $(\mu, \mu + 0.6745\sigma]$  and  $(\mu, \infty)$ . The corresponding  $\theta$  parameters are  $\theta_{36}, \theta_{23}, \theta_{12}$ . Let  $\Theta(x_J)$  be the collection of  $\theta_{js}$ s corresponding to

the sets containing  $x_J$ . There are  $J$  such parameters. Define a step function that serves as a weighting factor

$$r(x_J) = 2^J \prod_{\theta_{js} \in \Theta(x_J)} \theta_{js}. \quad (6)$$

For our  $x_3$  between the fifth and sixth octals,  $r(x_3) = 2^3 \theta_{36} \theta_{23} \theta_{12}$ . The density at stage  $J$  is just the product of the weighting factor and the original parametric density, that is,

$$f(x_J | \mu, \sigma^2, \theta_{js}, j = 1, \dots, J, s = 1, \dots, 2^j) = \psi(x_J) r(x_J), \quad (7)$$

where  $\psi$  denotes the normal density

$$\psi(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}.$$

Note that the  $\theta_{js}$ s only appear in the weight function  $r(\cdot)$ . Obviously,  $\psi(\cdot)$  can be replaced by the density for any other parametric family but corresponding changes in  $r(\cdot)$  must be made. This is a highly flexible model because the  $\theta_{js}$  parameters are numerous. There are  $2^{J+1} - 2$  of them, so for  $J = 6$  there are 126 and for  $J = 8$  there are 510. One often determines  $J$  as a function of the number of independent sampling units  $n$ . A good practical choice seems to be  $J \doteq \log_2(n)$ .

To perform a Bayesian analysis with these sampling distributions, we need a joint prior distribution on the  $\theta_{js}$  parameters. The  $\theta_{js}$ s are easily interpretable, so meaningful prior information may exist on many of them. An extreme case is choosing  $\theta_{11} = \theta_{12} = 0.5$  with prior probability 1. This gives prior probability 1 to the median being  $\mu$  for any  $J$ . Nonetheless, there are far too many parameters to choose a distribution that reflects meaningful prior information on all of the  $\theta_{js}$ s, so reference priors are also incorporated. Typically, meaningful prior information would be restricted to parameters from the first few stages  $j$ .

These  $\theta_{js}$  parameters are all probabilities, so it is convenient to use beta distributions. When the second subscript is odd, we assume for  $j = 1, \dots, J, k = 1, \dots, 2^{J-1}$  that

$$\theta_{j,2k-1} \sim \text{Beta}(\alpha_{j,2k-1}, \alpha_{j,2k})$$

with  $\theta_{j,2k} = 1 - \theta_{j,2k-1}$ . In alternative notation, the consecutive pairs have Dirichlet distributions,

$$(\theta_{j,2k-1}, \theta_{j,2k}) \sim \text{Dirich}(\alpha_{j,2k-1}, \alpha_{j,2k}).$$

We assume that all such pairs are independent in the prior.

For any parameters on which meaningful prior information is available, the  $\alpha_{js}$ s are chosen to reflect that information. However, in terms of defining a workable prior for all of the  $\theta_{js}$  parameters, we have not accomplished a great deal. We have reduced the problem of choosing a joint prior on  $2^{J+1} - 2$  parameters to choosing  $2^{J+1} - 2$  hyperparameters, the  $\alpha_{js}$ s. To make this more manageable, for all parameters without meaningful prior information we typically assume that  $\alpha_{js} = \alpha \rho(j)$  for some constant  $\alpha$  and nondecreasing function  $\rho(\cdot)$ . For fixed  $\rho(\cdot)$  the hyperparameter  $\alpha$  indicates strength of prior belief in the original parametric family. Often we take  $\rho(j) = j^2$ .

One advantage of the  $\alpha_{js} = \alpha \rho(j)$  priors is that on average they give the same probabilities as the original parametric distributions. Thus, if  $Y \sim N(\mu, \sigma^2)$ ,  $X_J$  has the stage  $J$  distribution, and  $A$  is any set,

$$E[\Pr(X_J \in A)] = \Pr(Y \in A)$$

where the expectation is over the prior on the  $\theta_{js}$  parameters. To see this, first observe that by our construction, if one fixed  $\theta_{js} = 0.5$  for all  $j$  and  $s$ , then clearly  $X_J \sim N(\mu, \sigma^2)$  and, in particular, the weight function (6) is  $r(x_j) = 1$ , so the result follows from (7). With these reference priors, for any  $x_J$ , the  $\theta_{js}$  parameters in  $r(x_J)$  are independent with mean 0.5, so  $E[r(x_J)] = 1$  and thus, again using (7), the average density is the normal density. Consequently, these reference priors are particularly

appropriate if we believe the data may come from the original parametric family but want to allow for other possibilities.

Correspondingly, prior or posterior distributions that focus high probability on regions around  $\theta_{js} = 0.5$  for all  $js$  will behave very much like normal distributions. This occurs whenever  $\alpha$  is large in  $\alpha_{js} = \alpha\rho(j)$ . Moreover, with  $\rho(j)$  increasing there is a strong prior tendency when  $j$  is large for a new level to behave like the previous level. This tendency can be overcome by data but this tendency also allows us to use numbers of parameters that are comparable to the number of observations in the data without “overfitting” the model.

On the other hand, when  $\alpha$  is small, the distribution is more “nonparametric.” Let  $A_J$  be a set in the  $J$ th level partition. When  $\alpha$  is small, an observation in  $A_J$  has a large effect on all the posterior beta distributions of  $\theta_{js}$ s associated with  $A_J$ , thus causing high probability for  $A_J$  in the posterior distribution. Since  $A_J$  is a set in the finest partition considered, this causes jagged, approximating discrete, behavior in the posterior.

The  $J$ th stage generalized sampling distribution, say  $G$ , depends on the  $\theta_{js}$ s and the original  $N(\mu, \sigma^2)$  distribution.  $G$  together with the reference prior on the  $\theta_{js}$ s determined by  $\alpha$  and  $\rho$  defines a random distribution that, because it is random, itself has a distribution called a finite Polya tree, which we write

$$G \sim PT_J(\alpha, \rho, N(\mu, \sigma^2)).$$

A prior on  $(\mu, \sigma)$  implies that the median  $\mu$ , the quartiles, octiles, and so on, are uncertain. This has the effect of smoothing out the abrupt jumps at these points that are noticeable in Figure . Figure d contains a realization of a third stage Polya tree that is conditional on  $\mu = \mu_0$ ,  $\sigma^2 = \sigma_0^2$ , and the  $\theta_{js}$ s. It also contains a realization of a mixture of a third stage Polya tree that is integrated over  $\mu$  and  $\sigma^2$ , but still conditional on the specified  $\theta_{js}$ s. Specifically,  $\mu \sim N(\mu_0, (\sigma/10)^2)$  and  $\sigma \sim N(\sigma_0, (\sigma_0/10)^2)$ .

The distribution on  $G$  obtained by randomly generating the  $\{\theta_{js}\}$ s according to the reference prior with  $\alpha_{js} = \alpha\rho(j)$ , but averaged over a prior on  $(\mu, \sigma)$  is called a *mixture of Polya trees (MPT)*. Write

$$G \sim \int PT_J(\alpha, \rho, N(\mu, \sigma^2)) p(\mu, \sigma^2) d\mu d\sigma^2.$$

Hanson (2006) shows that for typical priors, for example,  $(\mu, \tau) \sim N(\mu_0, V) \times \text{Gamma}(a, b)$ , that the random MPT density  $g(u) = dG(u)/du$  is smooth. Polya trees and other versions of mixtures of Polya trees do not necessarily have this property, see Barron et al. (1999), Paddock (1999), and Berger and Guglielmi (2001).

**EXAMPLE 15.1.5.** *Galaxy Data.* For the MPT model on the galaxy data we used  $J = 5$ ,  $\rho(j) = j^2$ , normal-gamma reference priors, and an additional prior on  $\alpha$ ,  $\alpha \sim \text{Gamma}(5, 1)$ . The predictive density is given in Figure 15.4. The corresponding LPML is  $-220.9$ , which is not very good compared to the mixtures of normals models. That is because the MPT smooths the tails of the distribution more than the mixtures of normals, see for example the MPT density in Figure 15.4 around 15 and around 30.

**EXAMPLE 15.1.6.** *Toenail Data.* These data were previously examined in Example 8.5.1. Recall that the model there was a *logistic regression* with *random effects* and the goal was to model the probability of moderate or severe toenail separation as a function of time and treatment. Random effects were assumed to be normally distributed and were used to model heterogeneity across individuals in the study and to model correlation among repeated observations on the same individual. Here, we replace the normality assumption on random effects with

$$\gamma_1, \dots, \gamma_{294} | G \stackrel{iid}{\sim} G,$$

$$G | \mu, \sigma^2 \sim PT_8(\alpha, j^2, N(\mu, \sigma^2)).$$

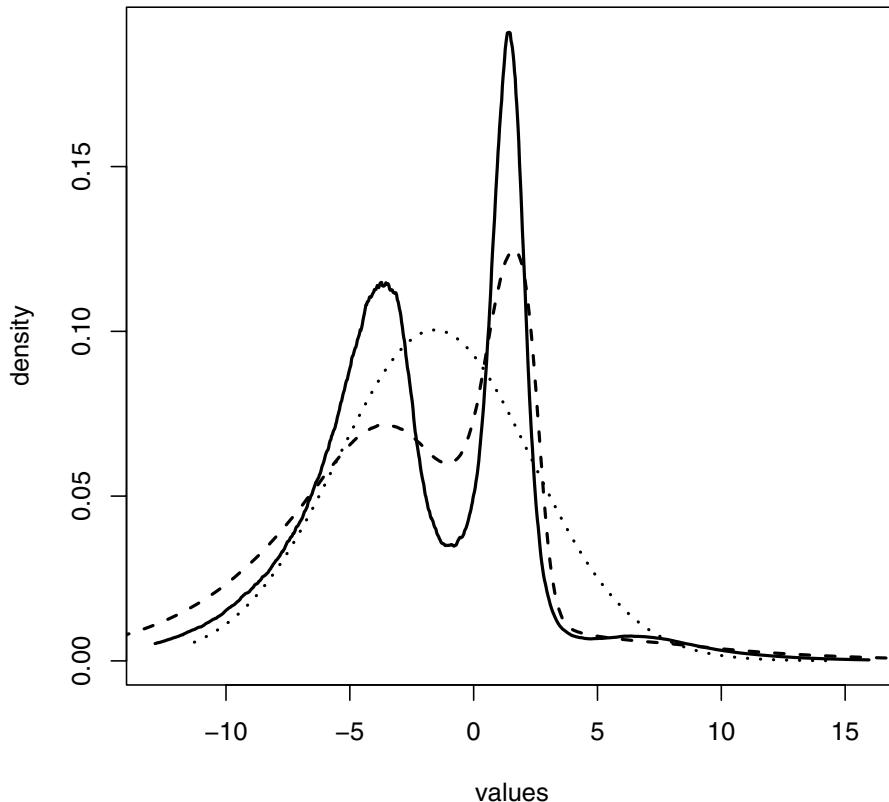


Figure 15.6: *Toenail data estimated random effects distribution under the MPT ( $\alpha = 0.1$ , solid line), MPT ( $\alpha = 1$ , dashed line), and normal ( $\alpha \rightarrow \infty$ , dotted line) models.*

Together with the prior on  $(\mu, \sigma^2)$  from Example 8.5.1, this constitutes an MPT prior on the  $\gamma_s$ . This methodology allows the evaluation of the normality assumption of the random effects and the implications of potential mis-specification. For example, it allows for multiple modes in the random effects distribution. Multiple modes suggest that there may be distinct subpopulations of individuals in the study whose probabilities of toenail separation differ. Different reference priors for the  $\theta_{js}$ s were considered using three values of  $\alpha$ ,  $\alpha = 0.1, 1, 10$ , to reflect increasing degrees of belief in normality for the random effects.

Figure compares the predictive distributions of a new random effect, say  $\gamma_{295}$ , not in the observed data, using the two best MPT models as determined by DIC and LPML (see Table 15.1) and the normal theory model. The plot clearly shows deviation from normality in such a way that the patients could be divided into two or three groups according to their resistance against infection and accompanying toenail separation. Note the smooth appearance of the MPT predictive distributions. The fact that the density of the generalized distributions contains jumps typically washes out in the posterior analysis of an MPT.

Table presents results from fitting the normal model and the three MPT models. The Polya trees outperform the normal model using either the LPML or DIC statistic, suggesting that the MPT model is better both for explaining the observed data and from a predictive viewpoint.

Table also shows the effects, in this instance, of incorrectly assuming a normal distribution for

Table 15.1: Posterior means, model comparison criteria, and 95% probability intervals for parameters of the toenail GLMM:  $\beta_1(\text{Trt})$ ,  $\beta_2(\text{Time})$ ,  $\beta_3(\text{Trt} \times \text{Time})$ .

	<b>Normal</b>	<b>MPT</b>		
		$\alpha = 10$	$\alpha = 1$	$\alpha = 0.1$
$\beta_1$	-0.159	-0.051	0.292	0.364
$\beta_2$	-0.393	-0.390	-0.393	-0.376
$\beta_3$	-0.138	-0.136	-0.130	-0.129
$\mu$	-1.604	-1.999	1.510	-0.181
$\sigma^2$	16.025	16.728	52.457	23.631
DIC	964.2	954.5	906.3	909.9
LPML	-484.0	-482.5	-465.3	-470.5
Posterior Intervals				
$\beta_1$	(-1.290, 0.966)	(-1.145, 1.090)	(-0.681, 1.186)	(-0.486, 1.198)
$\beta_2$	(-0.478, -0.305)	(-0.478, -0.304)	(-0.488, -0.307)	(-0.472, -0.289)
$\beta_3$	(-0.271, -0.005)	(-0.270, -0.003)	(-0.274, 0.010)	(-0.266, 0.007)
$\mu$	(-2.473, -0.790)	(-3.544, -0.479)	(-3.402, 3.884)	(-6.869, 5.664)
$\sigma^2$	(10.445, 22.053)	(9.562, 24.867)	(8.541, 116.273)	(1.436, 42.767)

the random intercepts  $\gamma_i$ . The intervals for  $\mu$  become substantially narrower as the random effects become more normal ( $\alpha$  increases). Comparing the interval estimates to 0, none of the models shows a statistically important baseline effect ( $\beta_1$ ) for treatments. This should be the case for a randomized experiment because the first measurements should be taken after treatment assignment but before treatments are actually applied or effective. All models show time effects ( $\beta_2$  different from 0) and that the treatment coded 1 works better over time ( $\beta_3 > 0$ ). The posterior probability of  $\beta_3 > 0$  was 2.03%, 2.18%, 3.56%, and 3.24% for  $\alpha = \infty$ ,  $\alpha = 10$ ,  $\alpha = 1$ , and  $\alpha = 0.1$ , respectively. Although all models show evidence of differential treatment ( $\beta_3$ ) effects, the two MPT models with weak normality assumptions show a reduction in the posterior evidence. Note that  $\alpha = 1$  is the best fitting model. It fits better than  $\alpha = 0.1$  which is “more nonparametric.”

Christensen, Hanson, and Jara (2008) give an example on Ache monkey hunting and discuss full conditional distributions.

**EXERCISE 15.6.** WinBUGS code for fitting Example 15.1.5 is available in `MPTdensity.odc` on the book website. The code includes computing the CPO statistics as well as the predictive density (and interval estimates) over a grid of points. Duplicate the mixture of Polya trees fit in Figure 15.4. Rerun the code with  $J = 6$ . Do the predictive density and LPML change much? Try fixing  $\alpha = 1$  and  $\alpha = 10$  (with  $J = 5$ ). How do the predictive densities and LPML statistics change? How well is the MCMC mixing?

**EXERCISE 15.7.** The function `PTdensity` from `DPpackage` fits a finite or *infinite mixture of Polya trees* to data using a reference prior on  $(\mu, \sigma)$ . Use the function to duplicate the mixture of Polya trees fit in Figure 15.4. Rerun the function with  $J = 6$  (called `M` in the function). Do the predictive density and LPML change much? Try fixing  $\alpha = 1$  and  $\alpha = 10$  (with  $J = 5$ ). How do the predictive densities and LPML statistics change? How do the normal parameters  $(\mu, \sigma)$  mix as  $\alpha$  gets smaller? Code is in the file `Chap15DPpackage.txt`.

## 15.2 Flexible Regression Functions

The nonparametric regression problem is typically cast as estimating the mean function  $m(\cdot)$  from data  $\{(x'_i, y_i)\}_{i=1}^n$  in the model

$$y_i = m(x_i) + \varepsilon_i, \quad E(\varepsilon_i) = 0,$$

where the  $\varepsilon_i$ s are iid and the  $x_i$ s are treated as fixed. In some applications the shape of the  $\varepsilon_i$  distribution is of interest as well. We initially assume  $x_i$  is univariate but later discuss the case when  $x_i$  is a vector of predictors.

One approach to this problem, the only one we consider, relies on the fact that when  $m$  is smooth and defined on a closed bounded set, it can be represented as a linear combination of basis functions. In other words, given basis functions  $\{\phi_k\}_{k=1}^{\infty}$ , we can write

$$m(x) = \sum_{k=1}^{\infty} \beta_k \phi_k(x).$$

If the basis functions are orthonormal in the sense of having

$$\int \phi_k^2(x) dx = 1; \quad \int \phi_j(x) \phi_k(x) dx = 0, \quad j \neq k,$$

then it is easily seen that for  $m$  with  $\int [m(x)]^2 dx < \infty$ ,

$$\beta_k = \int m(x) \phi_k(x) dx.$$

Orthonormal bases make certain common mathematical calculations trivial, but are not required for this approach. Moreover, in practice it is not the integral properties of the functions that are important but rather their properties when evaluated at the finite number of points in our data. Popular choices for  $\{\phi_k\}$  are polynomials, the Fourier series (sines and cosines), wavelet bases, spline bases, and B-spline bases. Figure 15.7 presents two cosine basis functions from among  $\{\cos[x(k-1)\pi]\}_{k=1}^{\infty}$  and three Haar wavelet basis functions. Standard basis functions are usually defined on the interval  $[0, 1]$  so, in applications, the predictor variables usually need to be rescaled before they are used.

It is impossible to estimate  $\{\beta_j\}_{j=1}^{\infty}$  with finite data. Instead we must rely on an approximation

$$m(x) \doteq \sum_{k=1}^K \beta_k \phi_k(x).$$

The basis functions are typically ordered in some fashion from broad functions that indicate a rough trend to functions that model detailed local behavior. Typically,  $\phi_1(x) \equiv 1$ . If one fixes  $K$  and assumes iid normal errors then a standard linear model is obtained:

$$y_i = \beta_1 + \beta_2 \phi_2(x_i) + \cdots + \beta_K \phi_K(x_i) + \varepsilon_i, \quad \varepsilon_i \stackrel{iid}{\sim} N(0, 1/\tau)$$

with

$$X = \begin{bmatrix} 1 & \phi_2(x_1) & \cdots & \phi_K(x_1) \\ 1 & \phi_2(x_2) & \cdots & \phi_K(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \phi_2(x_n) & \cdots & \phi_K(x_n) \end{bmatrix}$$

in the characterization  $Y = X\beta + e$  of Chapter 9.

**EXAMPLE 15.2.1.** Brinkman (1981) presents data on the amount of nitric oxide and nitric dioxide in the exhaust of a single-cylinder test engine fueled by ethanol. The response is in  $\mu\text{gs}$  (micrograms) per joule and the predictor is a measure of the air-to-fuel ratio, called the equivalence ratio. The data are part of a larger set used throughout MathSoft (1999) to illustrate various smoothing techniques and are available on our website. We restrict the predictor values  $\tilde{x}$  to the domain  $0.5 \leq \tilde{x} \leq 1.3$  and rescale  $\tilde{x}$  into  $x$  so that  $0.0 \leq x \equiv (\tilde{x} - 0.5)/0.8 \leq 1$ . We considered two sets of basis functions both with  $K = 5$ : the cosine,  $\phi_k(x) = \cos[x(k-1)\pi]$ , and Legendre polynomial bases. Legendre polynomials are simply a set of orthogonal polynomials. (They can be obtained

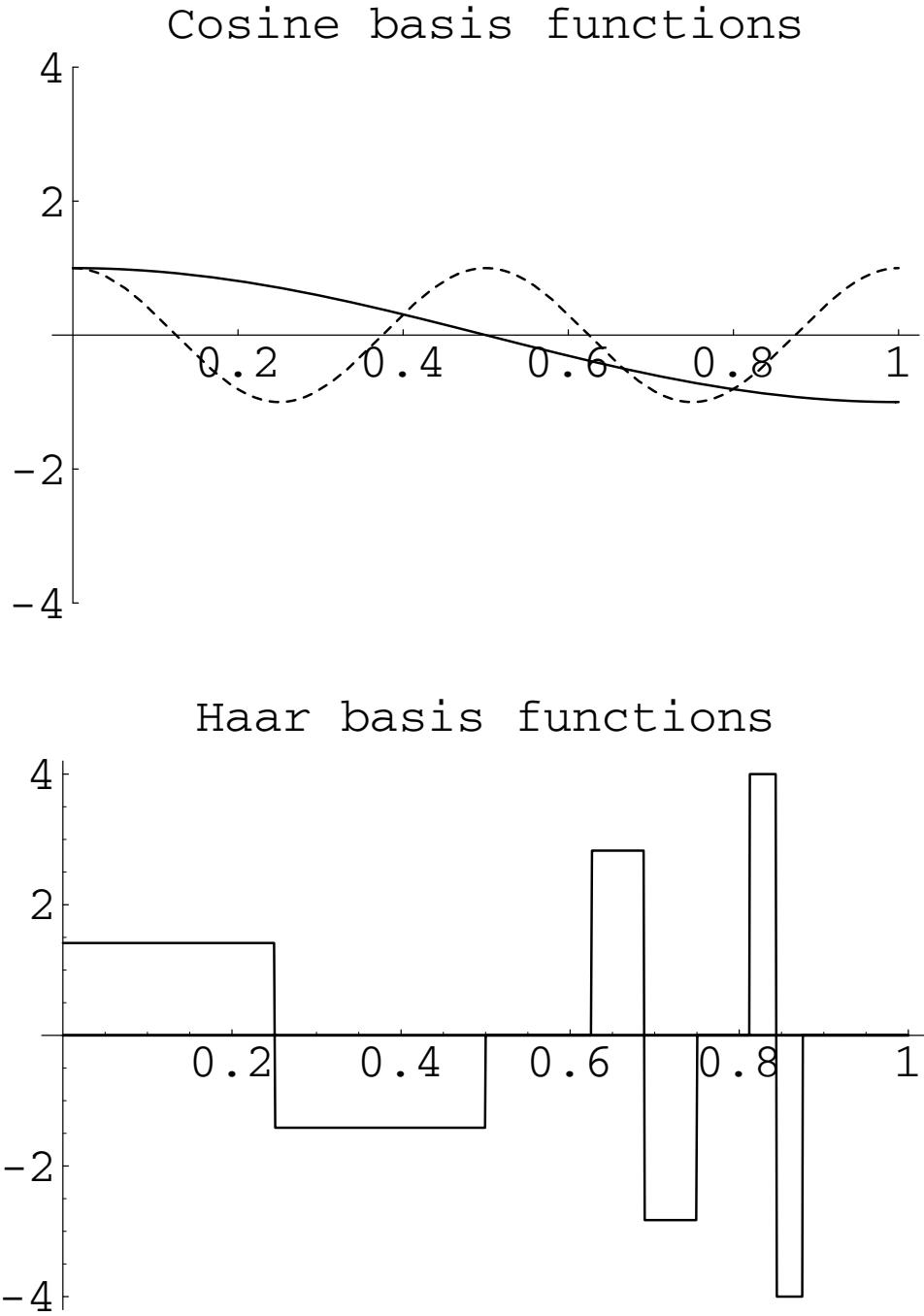


Figure 15.7: Top: Cosine basis functions  $\cos(x\pi)$  (solid) and  $\cos(x4\pi)$  (dashed). Bottom: Three Haar basis functions with supports  $[0, 1/2]$ ,  $[5/2^3, 6/2^3]$ ,  $[13/2^4, 14/2^4]$ , i.e.,  $\phi_{2,1}$ ,  $\phi_{4,6}$ , and  $\phi_{5,14}$ .

by applying the Gram-Schmidt procedure to the sequence of polynomials  $1, x, x^2, \dots$  defined on the unit interval.) Independent  $N(0, 1000)$  priors were placed on the regression coefficients independent of the precision prior  $\tau \sim \text{Gamma}(0.001, 0.001)$ . The posterior mean of the regression function and 95% probability intervals are given in Figure 15.8 for  $\tilde{x}$  in the domain. Trying to extrapolate results

beyond the range of the data is always dangerous but for these data the figure illustrates that it would be particularly dangerous using the polynomial model. Informative priors could be incorporated as discussed in Chapter 9 but typically  $K$  is large enough that only partial prior information would be practical. The choice of basis functions, domain of  $\tilde{x}$ , and  $K$  all affect posterior inference.

When fitting models that use basis functions defined on  $[0, 1]$ , the predictor variable  $\tilde{x}$  must be rescaled to  $x$  on  $[0, 1]$ . However, in our plots the horizontal axis corresponds to the original scale of the predictor. It is a simple matter to transform any inference about  $m(x)$  such as posterior means, medians, or intervals into an inference about the corresponding regression function for  $\tilde{x}$ . For example, with original data  $\{(\tilde{x}_i, y_i) : i = 1, \dots, n\}$  one often obtains fitted values  $\hat{y}_i = \hat{m}(x_i)$ . The appropriate plot of the fitted values on the original scale is obtained from  $\{(\tilde{x}_i, \hat{y}_i) : i = 1, \dots, n\}$ . When plotting a function  $q$  over a grid, one need only keep track of the  $\tilde{x}$  value that corresponds to each point  $(x, q(x))$  and plot  $(\tilde{x}, q(x))$ .

Using too many basis functions can be a problem. It is well known that an  $(n - 1)$  degree polynomial fits data  $\{(x_i, y_i)\}_{i=1}^n$  perfectly but that the functions can do very bizarre things at  $x$  values between those in the data. This constitutes an example of overfitting, i.e., including too many parameters for the number of observations. Other basis functions like sines and cosines, that are positive almost everywhere on the unit interval, can display many of the more bizarre features encountered when overfitting polynomials. Two approaches to dealing with overfitting are using the data to determine a reasonable value of  $K$  or letting  $K$  be large but using the data to essentially eliminate individual basis functions. Reasonable values of  $K$  are often chosen using model selection criteria. Frequentists often use one equivalent to the  $C_p$  statistic, see Christensen (2001a, Section 7.4). Eliminating, or at least deemphasizing, individual basis functions is known as *thresholding*.

Thresholding requires that substantial data-driven evidence exists for its importance before allowing a basis function into the model. Typically more evidence is required for larger  $k$ . One simple approach is to put a prior distribution on each  $\beta_k$  that incorporates positive probability that the parameter is 0. This approach, advocated by Smith and Kohn (1996), can be formulated as writing

$$\beta_k \equiv \gamma_k \beta_k^*$$

where  $\gamma_k \sim \text{Bern}(q_k)$ . The  $\beta_k^*$ 's have continuous distributions and everything is assumed independent. The  $q_k$ s are known and are typically decreasing towards 0. For moderate  $K$  this can be programmed in WinBUGS. More generally, Bayesian thresholding places mixture priors on basis coefficients that give positive probability to some coefficients being very small or zero. Clyde and George (2004) discuss priors of this type in more detail.

**EXAMPLE 15.2.2.** For the ethanol data using the cosine basis with  $K = 10$  we consider the rather naive prior  $\gamma_k$  iid  $\text{Bern}(0.5)$ . Heretically, in this example we use a prior on  $\beta_j^*$  determined by the data. (Every congregation of four or more contains a heretic.) Using least squares to fit the linear model with all basis functions up to  $K$ , let  $b_k$  be the least squares estimate of  $\beta_k$ . This has variance  $\sigma^2 v_k$ . The data based prior has  $\beta_k^* | \sigma^2 \sim N(b_j, 10\sigma^2 v_j)$  and a reference prior on the precision. Figure 15.9 shows the estimate of the regression function on the  $\tilde{x}$  scale with 95% probability intervals. Five of the 10 basis functions have posterior probability  $\Pr(\gamma_j = 0 | y)$  less than the prior value of 0.5.

A popular Bayesian alternative to fixing the number of components when fitting basis function models is to place a prior on  $K$  and implement the reversible jump algorithm of Green (1995). *Reversible jump MCMC* approximates posterior inference over a model space where each model has a parameter vector of possibly different dimension. A prior probability is placed on each of  $K = 1, 2, \dots, K_0$ , where  $K_0$  is some natural upper bound chosen such that consideration of  $K > K_0$  would be superfluous. Reversible jump for the regression problem (in the context of a spline basis) is discussed in Denison et al. (2002) and used, for example, by Mallick et al. (1999) and Holmes and Mallick (2001).

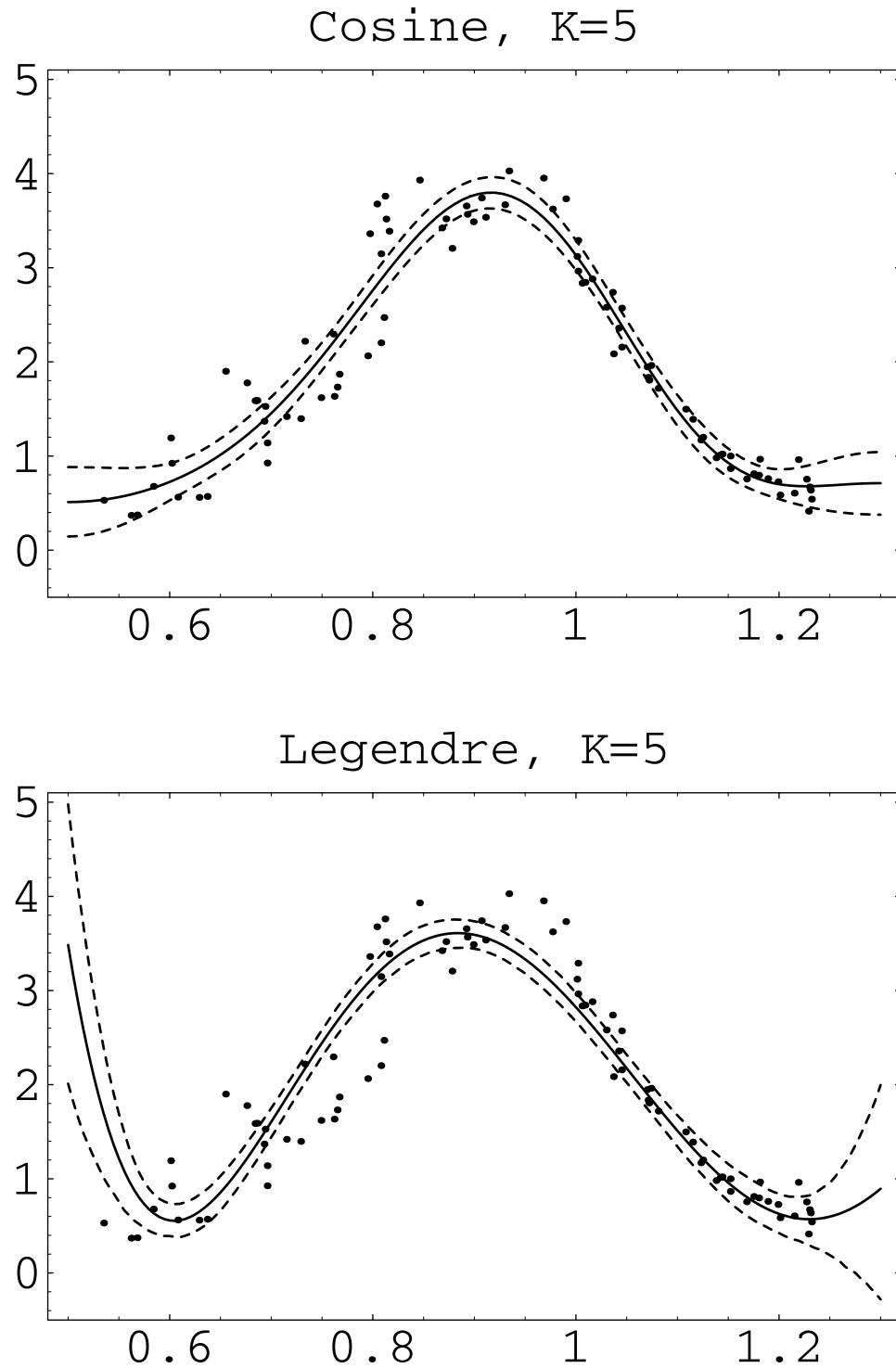


Figure 15.8: Ethanol data: Mean function estimate using cosine and Legendre bases,  $K = 5$ .

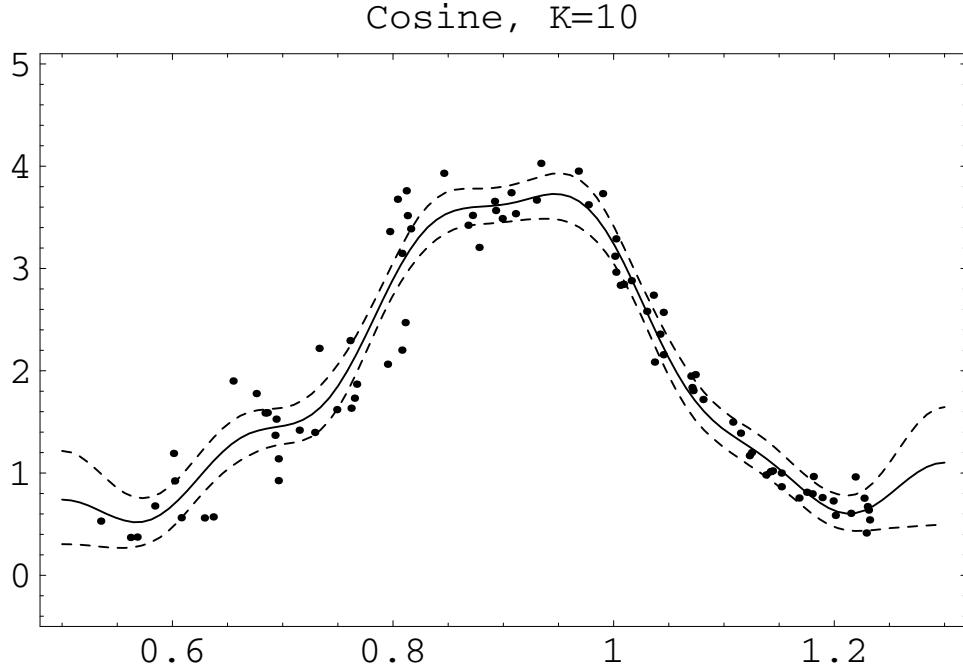


Figure 15.9: Ethanol data: Estimates of regression mean functions using a cosine basis and thresholding.

An unusual class of bases are wavelet bases. Wavelets are useful for modeling functions whose behavior changes dramatically at different locations. Such functions are sometimes called “spatially inhomogeneous.” Think of a topographical map of the western United States. Much of the map will have relatively flat homogeneous areas but at the edges of mountains the scale changes abruptly. Wavelets can capture such phenomena and so are used extensively in image processing. A nice, short introduction to Bayesian wavelets and thresholding is Vidakovic (1998). Müller and Vidakovic (1999) discuss Bayesian wavelet modeling in detail.

The key feature of wavelets is that they are 0 except over an increasingly smaller range of  $x$  values. The simplest wavelet basis was developed by Haar (1910). On the interval  $[0, 1]$  fitting the Haar basis to level  $k - 1$  is equivalent to fitting a step function in which each step has length  $1/2^k$ , that is, each step is a multiple of an indicator function  $I_{(\{j-1\}/2^k, j/2^k]}(x)$ ,  $j = 1, \dots, 2^k$ . Rather than using increasingly smaller indicator functions, the Haar wavelet basis combines indicator functions so that the individual basis functions will be orthonormal in the sense defined earlier.

The Haar wavelet basis (as well as other wavelet bases) is conveniently enumerated using a double index. The basis functions can be derived from a *mother wavelet*  $\psi$  that for the Haar basis is defined as

$$\psi(x) = \begin{cases} 1 & 0 \leq x < 0.5 \\ -1 & 0.5 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}.$$

Individual wavelet basis functions are defined through  $\phi_{kj}(x) = \psi(2^k x - j)2^{k/2}$  for  $k = 0, \dots, \infty$ , and  $j = 0, \dots, 2^k - 1$ . Thus  $\phi_{kj}(x)$  is nonzero only on the interval  $[j/2^k, (j+1)/2^k]$ . Figure 15.7 showed three of the Haar basis functions:  $\phi_{1,1}$ ,  $\phi_{3,6}$ , and  $\phi_{4,14}$ . The orthonormal basis consists of the Haar wavelets together with the *father wavelet*  $I_{[0,1]}(x)$ . With the father corresponding to the

intercept, the wavelet basis form for the regression function is

$$m(x) = \beta_0 + \sum_{k=0}^{\infty} \sum_{j=0}^{2^k-1} \beta_{kj} \phi_{kj}(x).$$

In practice,  $k = 0, \dots, K$ . In this notation  $k$  indexes the scale of the basis function whereas  $j$  determines location. The most important ideas in the definition of wavelets are that the support of the mother function is bounded and that the subsequent (children) functions are defined as indicated by reductions in scale and by translations. Having orthogonal  $\phi_{kj}$  functions is of little importance.

It seems that just about any function that is 0 off of the unit interval can be used as a mother wavelet  $\psi$ . Some commonly assumed properties are that  $\psi$  is continuous and that it integrates to 0 over the unit interval. Many commonly used mother wavelets are continuous but cannot be written in closed form. We present a few continuous wavelets that can be written down. These functions might be adjusted so that they integrate to 0 on the unit interval (not important) but they need to be defined so that they are (for practical purposes) 0 outside the unit interval. Shannon's mother is

$$\psi(x) = \frac{\sin(2\pi x) - \sin(\pi x)}{\pi x} I_{[0,1]}(x).$$

A wavelet mother shaped like a Mexican hat is given by

$$\psi(x; \sigma) = \frac{1}{\sqrt{2\pi}\sigma^3} \left( 1 - \frac{(x-0.5)^2}{\sigma^2} \right) \exp[-(x-0.5)^2/2\sigma^2].$$

This is essentially the negative second derivative of a normal density and depends on a scaling hyperparameter  $\sigma$ . The Mexican hat can also be approximated by a difference of Gaussian (normal) densities using two scaling parameters

$$\psi(x; \sigma_1, \sigma_2) = \frac{1}{\sigma_1 \sqrt{2\pi}} \exp\left(-\frac{(x-0.5)^2}{2\sigma_1^2}\right) - \frac{1}{\sigma_2 \sqrt{2\pi}} \exp\left(-\frac{(x-0.05)^2}{2\sigma_2^2}\right).$$

Finally, a wavelet mother based on the beta distribution is

$$\psi(x|\alpha, \beta) \propto \left[ \frac{\beta-1}{1-x} - \frac{\alpha-1}{x} \right] x^{\alpha-1} \cdot (1-x)^{\beta-1} I_{[0,1]}(x).$$

For large  $k$ , wavelet basis functions can model very localized behavior. In Figure 15.7, contrast the Haar basis functions to the cosine basis functions that oscillate over the entire region. Wavelets can model highly inhomogeneous functions but also require extra care to ensure that mean estimates do not follow the data too closely, i.e., avoid overfitting. For moderate  $K$  you might fit polynomials or trig functions without thresholding but you probably don't want to fit wavelets without thresholding.

**EXAMPLE 15.2.3.** We fit the Haar wavelet model to the ethanol data obtaining a step function with steps of length 1/16 on the  $x$  scale. The model is

$$y_i = \beta_0 + \sum_{k=0}^3 \sum_{j=1}^{2^k} \beta_{kj} \phi_{kj}(x_i) + \varepsilon_i, \quad \varepsilon_i \stackrel{iid}{\sim} N(0, \sigma^2).$$

To incorporate thresholding, we introduce 0-1 parameters  $\gamma_{kj}$  that determine whether  $\phi_{kj}$  is included into the model

$$y_i = \beta_0 + \sum_{k=0}^3 \sum_{j=1}^{2^k} \gamma_{kj} \beta_{kj}^* \phi_{kj}(x_i) + \varepsilon_i, \quad \varepsilon_i \stackrel{iid}{\sim} N(0, \sigma^2).$$

We use independent  $\text{Bern}(1/2^k)$  priors on the  $\gamma_{kj}$ s, so that it becomes progressively more difficult

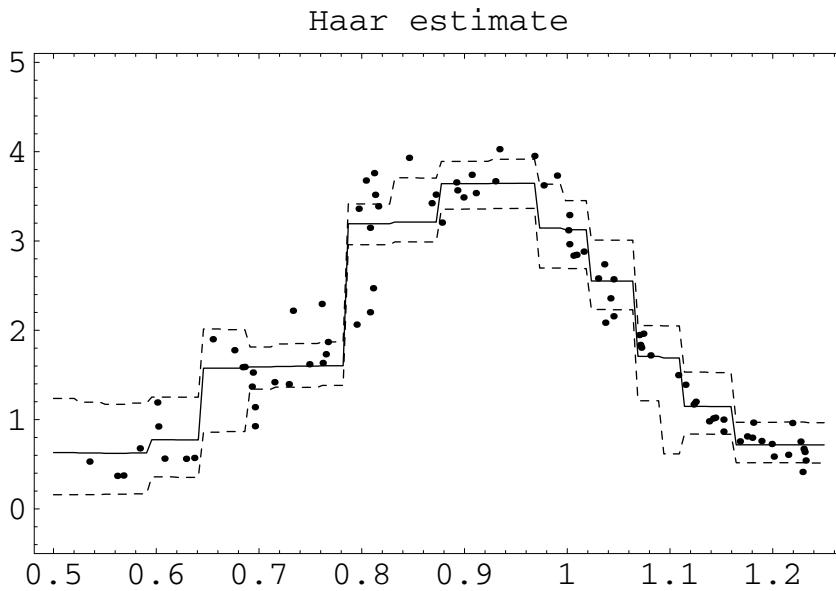


Figure 15.10: Ethanol data: Estimates of regression mean functions using Harr wavelets.

to include higher order  $\phi_{kj}$  functions. We use independent prior distributions,  $N(0, 1000)$  on the  $\beta_{kj}^*$ s and  $\text{Gamma}(0.001, 0.001)$  on the precision. Figure 15.10 shows the resultant mean function estimate on the  $\tilde{x}$  scale with 95% probability intervals. Four of the 16 basis functions had posterior probabilities less than 0.1 of being included in the model.

Another approach to nonparametric regression is the use of splines. The fundamental idea of using splines is simply connecting the dots in an  $x, y$  plot. Suppose we have data  $(x_i, y_i), i = 1, \dots, n$  in which the  $x_i$ s are ordered from smallest to largest. Using linear splines to fit a regression function simply fits a line segment between the consecutive pairs of points. Cubic splines fit a cubic polynomial between every pair of points rather than a line. The reason for using cubic splines is to make the curve look smooth. Although there is only one line you can fit between two points, there are many cubic polynomials. You pick the cubic polynomials so that the overall curve connecting all the data points has continuous first and second derivatives. These conditions determine what each individual cubic polynomial must be. Note that all of the action here has to do with what the function looks like between the data points. All of the data points are fitted perfectly.

**EXERCISE 15.8.** Find a cubic spline function for connecting the points  $(0, 0)$ ,  $(1, 1/3)$ ,  $(2, 4/3)$ ,  $(3, 1/3)$ , and  $(4, 0)$ . Hint: The first function is  $(x^3/3)I_{[0,1]}(x)$ . The value, first, and second derivatives of the second function must agree with those of the first function at  $x = 1$  and the second function is maximized at  $x = 2$  providing four equations for the four unknown parameters of the second cubic polynomial.

Of course, subtleties are often added. Instead of connecting the dots at the data points, the polynomials can be connected at other points called *knots*. Many spline bases are built from truncated polynomials. For example  $\{(x - a_j)_+^3\}_{j=1}^J$  is a subset of a cubic spline basis where  $\{a_j\}_{j=1}^J$  are the knots and  $(x)_+$  is equal to  $x$  when  $x > 0$  and equal to 0 otherwise. With splines, it is not crucial to rescale the predictor between 0 and 1. Also, regression coefficients can be penalized, that is, shrunk towards some predetermined value (often 0) to give imperfect fitting of the data but presumably superior predictive ability. Penalties exist to avoid overfitting. Priors serve the same purpose

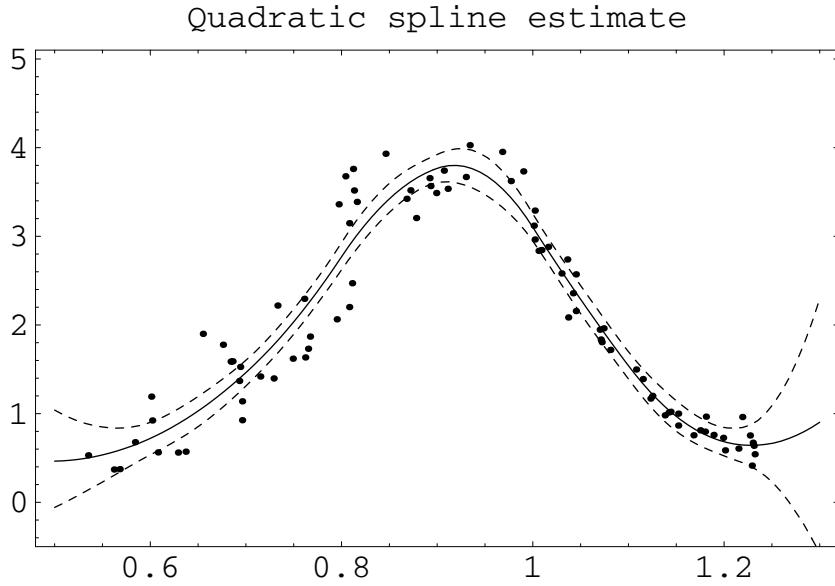


Figure 15.11: Ethanol data: Estimate of mean regression function using quadratic splines.

as penalizing regression coefficients because they shrink coefficients towards the center of the prior distribution. The extent of the penalty relates to the prior variability.

**EXAMPLE 15.2.4.** Crainiceanu et al. (2004) outline a strategy for fitting spline models in WinBUGS. We apply their approach by fitting a quadratic spline model to the ethanol data. Specifically, the model is

$$y_i = \beta_0 + \beta_1 \tilde{x}_i + \beta_2 \tilde{x}_i^2 + \sum_{k=1}^9 b_k (\tilde{x}_i - a_k)_+^2 + \varepsilon_i,$$

where  $b_k | \sigma_b \stackrel{iid}{\sim} N(0, \sigma_b^2)$  independent of  $\varepsilon_i | \sigma_\varepsilon \stackrel{iid}{\sim} N(0, \sigma_\varepsilon^2)$ . Here, the knots  $\{a_k\}_{k=1}^9$  are defined as  $a_i = 0.4 + 0.1i$ , evenly spaced over the range of the predictor variable. Figure 15.11 gives the spline estimate along with 95% pointwise probability intervals.

Another popular approach is the use of *basis splines* or *B-splines*, see Lang and Brezger (2004) or Jara et al. (2009) for Bayesian treatments. Although B-splines borrows an idea from splines, it is more similar to using wavelets in that the individual basis functions are 0 outside of a small interval. The “mother” function for a B-spline basis of degree 2 is nonzero over  $[0, 3]$  and defined as

$$\psi(x) = \frac{x^2}{2} I_{[0,1]}(x) - \{[x - 1.5]^2 - 0.75\} I_{[1,2]}(x) + \frac{[3-x]^2}{2} I_{[2,3]}(x).$$

This is a bell-shaped curve, similar to a normal density centered at 1.5, but it is 0 outside the interval  $[0, 3]$  while still being smooth in that it is differentiable everywhere. The “spline” in B-spline is because  $\psi$  is a quadratic spline function, i.e., quadratics have been pasted together as a smooth function. A B-spline basis mother function  $\psi$  of degree  $d$  splices together  $(d+1)$  different  $d$ -degree polynomials over a finite interval so that the whole function is differentiable  $d-1$  times and looks like a mean-shifted Gaussian density, but nonzero only on a finite interval. Commonly  $d$  is either 2 or 3.

Although  $\psi$  could be used as a perfectly reasonable mother function for fitting wavelets, typically B-splines are used a bit differently. Nonetheless, the basis functions  $\phi_k$  are simply scale reduced and location-shifted versions of  $\psi(x)$ .

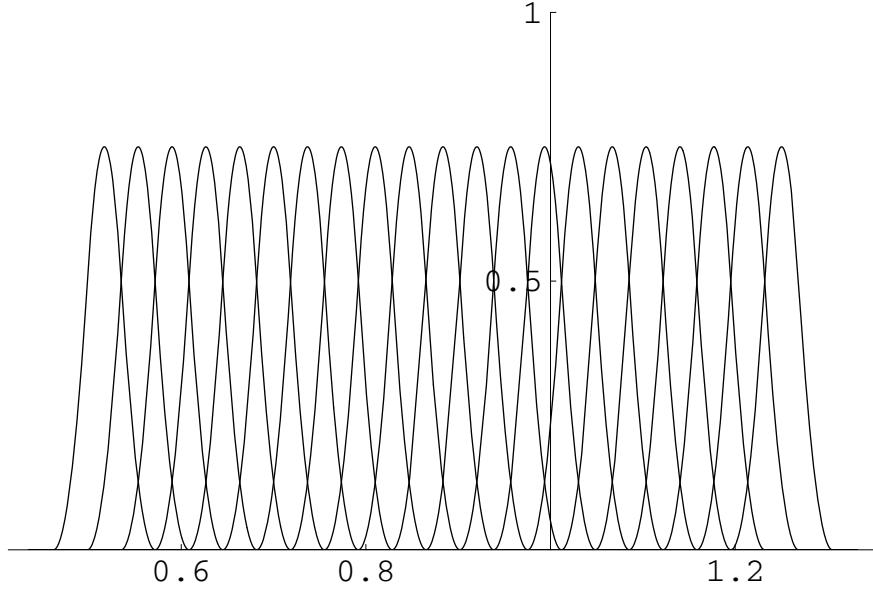


Figure 15.12:  $K = 21$  quadratic B-spline basis functions  $\{\phi_1(\cdot), \dots, \phi_{21}(\cdot)\}$  (from left to right) equispaced over  $[0.535, 1.232]$ .

**EXAMPLE 15.2.5.** Figure shows  $K = 21$  quadratic B-spline basis functions equally spaced over the range of the equivalence ratio predictor  $x$  for the ethanol data.

If  $[x_l, x_u]$  is the range of the observed predictor variable, the basis functions  $\phi_k(x)$  are defined so that they are centered at equally spaced locations, they have substantial overlap with one another, and they “bleed” over the ends of the interval  $[x_l, x_u]$ . There are many ways to do this and it does not seem to matter very much how you do it or even if you use a B-spline  $\psi$  function, as opposed to, say, a normal density. However, we employ the traditional B-spline approach.

The idea is best illustrated with an example. Suppose we want to use  $K = 5$  quadratic basis functions to cover the unit interval, i.e.,  $[0, 1] = [x_l, x_u]$ . Divide the unit interval into  $K - 2 = 3$  subintervals:  $[0, 1/3], [1/3, 2/3], [2/3, 1]$ . Determining these subintervals is the only place that  $K$  enters the process. Add two more comparable intervals onto each end of these so we have

$$[-2/3, -1/3], [-1/3, 0], [0, 1/3], [1/3, 2/3], [2/3, 1], [1, 4/3], [4/3, 5/3].$$

Just as the  $\psi$  function is defined over three subintervals of  $[0, 3]$ , rescale and translate  $\psi$  to give a first basis function  $\phi_1$  that covers the three intervals  $[-2/3, -1/3], [-1/3, 0], [0, 1/3]$ . Similarly,  $\phi_2$  covers the three subintervals contained in  $[-1/3, 2/3], \phi_3$  covers  $[0, 1], \phi_4$  covers  $[1/3, 4/3]$ , and  $\phi_5$  covers  $[2/3, 5/3]$ . For any  $K$ , each of the subintervals of our target interval  $[0, 1]$  has three nonzero  $\phi_k$  functions defined over it. For  $K$  quadratic basis functions on an arbitrary interval  $[x_l, x_u]$  this procedure becomes

$$\phi_k(x) = \psi \left\{ \frac{x - x_l}{\Delta} + 3 - k \right\}, \quad (1)$$

where  $\Delta = (x_u - x_l)/(K - 2)$  and  $k = 1, \dots, K$ . The B-spline regression function is taken as

$$m(x) \doteq \beta_0 + \sum_{k=1}^K \beta_k \phi_k(x). \quad (2)$$

EXERCISE 15.9. For  $d = 3$ , the cubic spline mother function is

$$\begin{aligned}\psi(x) &= \frac{x^3}{3}I_{[0,1]}(x) + \left\{-x^3 + 4x^2 - 4x + \frac{4}{3}\right\}I_{[1,2]}(x) \\ &\quad + \left\{-[4-x]^3 + 4[4-x]^2 - 4[4-x] + \frac{4}{3}\right\}I_{[2,3]}(x) + \frac{[4-x]^3}{3}I_{[3,4]}(x).\end{aligned}$$

For  $K = 5$  over  $[x_l, x_u] = [0, 1]$ , find the cubic B-spline basis functions. Can you find a formula for arbitrary  $K$  and  $[x_l, x_u]$  similar to equation (1)? How does  $\psi$  relate to your answer to Exercise 15.8?

The key characteristic of B-splines is that, similar to wavelets, the basis functions are 0 off of intervals that become increasingly small. Technically, for B-splines to constitute a basis function approach, we should be able to write the regression function exactly as an infinite linear combination of basis functions. To do that we would need to use a double index for the basis functions like we did for wavelets. The definition of  $\phi_k$  in (1) actually depends on  $K$  as well. With

$$\phi_{Kk}(x) \equiv \phi_k(x),$$

we could write

$$m(x) = \beta_0 + \sum_{K=1}^{\infty} \sum_{k=1}^K \beta_{Kk} \phi_{Kk}(x). \quad (3)$$

Using a finite approximation to this, as with our other basis functions, the higher order terms ( $K$  large) are less “smooth,” so there is a premium on shrinking the corresponding regression coefficients towards 0 to smooth any fitted model.

In practice, model (2) is used rather than an approximation to (3), so shrinking coefficients towards 0 seems inappropriate. All of the functions in (2) have the same level of smoothness. To ensure smoothness in the fitted model, we instead cause the regression coefficients to be similar for similar values of  $k$ .

One way to achieve smoothness is by using a first-order *random walk prior*. Specifically, with an informative or reference prior on  $\beta_0$ , let

$$\beta_k | \beta_0, \beta_1, \dots, \beta_{k-1}, \lambda \sim N(\beta_{k-1}, 1/\lambda), \quad k = 1, 2, \dots, K.$$

Alternatively, we could have standard priors on  $\beta_0$  and  $\beta_1$  with the random walk starting at  $k = 2$ . The larger  $\lambda$  is, the smaller the jumps can be between neighboring basis functions, and therefore  $m(x)$  in (2) is smoother. The parameter  $\lambda$  is often called a penalty term and has a frequentist interpretation related to fitting B-spline models through penalized likelihoods (Eilers and Marx, 1996). It is possible to elicit reasonable prior information on how “jumpy” the trends are, or one could employ a reference prior on  $\lambda$ . A reference prior allows  $\lambda$  to reflect an overall level of smoothness in the regression, but can yield quite bumpy fits if there are marked jumps in the data.

With the regression model defined by (2),  $\beta_0$  is nonidentifiable if and only if for every  $x_i$  there exist real numbers  $\alpha_k$  such that  $\sum_{k=1}^K \alpha_k \phi_k(x_i) = 1$ , see Christensen (2002, Proposition 2.1.6). In practice, this is extremely unlikely to happen unless the  $\phi_k$ s have been defined so that there exist  $\alpha_k$ s with  $\sum_{k=1}^K \alpha_k \phi_k(x) = 1$  for all  $x \in [x_l, x_u]$ . The basis functions in (1) have this property, see Exercise 15.10. In fact, for any way of defining basis functions that satisfy the criteria of small support, overlapping functions, and bleeding over the ends, it is quite likely that there exist  $\alpha_k$ s with  $\sum_{k=1}^K \alpha_k \phi_k(x_i) \doteq 1$ , causing a severe collinearity problem. A likely symptom of this is horrendous mixing problems within the MCMC. Both the identifiability and the numerical difficulty can be alleviated by imposing a side condition. The condition  $\sum_{k=1}^K \beta_k = 0$  is often used but, theoretically, dropping the intercept or any one of the  $\phi_k$  functions should also work. None of these side conditions should have any effect on the fitted regression function (except through using different priors).

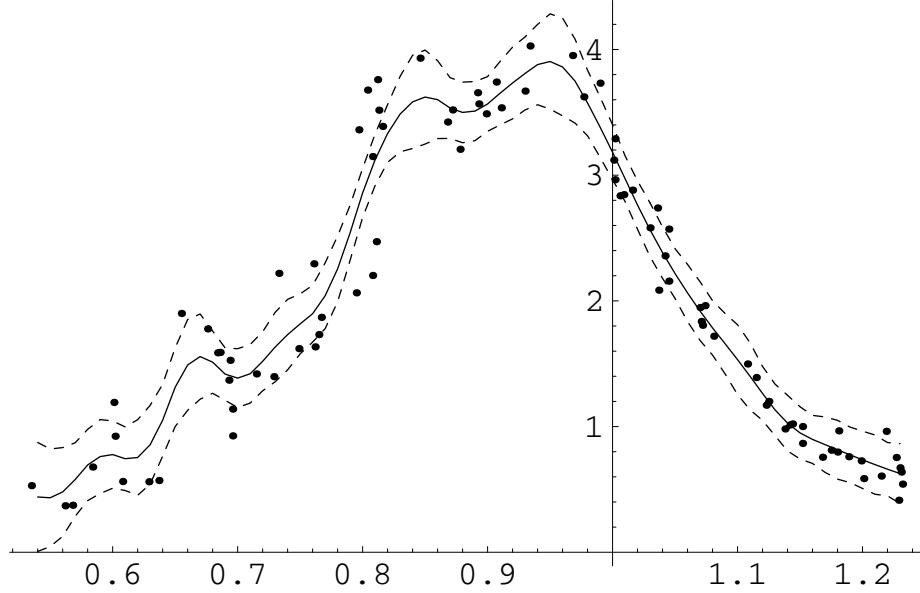


Figure 15.13: Estimated trend using quadratic B-splines with  $K = 21$  knots.

**EXERCISE 15.10.** For the basis functions in (1), show that  $\sum_{k=1}^K \phi_k(x_i)$  is a constant. Hint: Recall that exactly three nonzero basis functions overlap on every subinterval and argue that it is enough to show that for  $x \in [0, 1]$  the function  $x^2/2 - [(1+x) - 1.5]^2 + [3 - (2+x)]^2/2$  is a constant.

Eilers and Marx (1996), citing de Boor (1977), indicate that when using the basis functions in (1) there also exist different sets of  $\alpha_k$ s that for all  $x \in [x_l, x_u]$  make  $\sum_{k=1}^K \alpha_k \phi_k(x) = x$  and make  $\sum_{k=1}^K \alpha_k \phi_k(x) = x^2$ . This implies that the quadratic B-spline basis can fit second degree polynomials. Using a B-spline  $\psi$  function of degree  $d$ ,  $d$  is also the degree of the polynomial that can be fitted.

**EXAMPLE 15.2.6. Ethanol Data.** The model

$$y_i = \beta_0 + \sum_{k=1}^{21} \beta_k \phi_k(x_i) + \varepsilon_i$$

was fitted to the ethanol data with a random walk prior determined by

$$\beta_0 \sim N(0, 100000) \quad \perp\!\!\!\perp \quad \lambda, \tau \stackrel{iid}{\sim} \text{Gamma}(0.0001, 0.0001).$$

The posterior mean regression function is given in Figure 15.13 along with a 95% interval. Here  $[x_l, x_u] = [0.535, 1.232]$ .

Multivariate predictors  $x_i = (x_{i1}, \dots, x_{ir})'$  can be accommodated into series expansions by considering products of univariate basis functions. For example, in two dimensions with  $x = (x_1, x_2)'$ , simple products are formed as  $\phi_{jk}(x_1, x_2) = \phi_j(x_1)\phi_k(x_2)$ . The regression model is then

$$y_i = \sum_{j=1}^{K_1} \sum_{k=1}^{K_2} \beta_{jk} \phi_{jk}(x_{i1}, x_{i2}) + \varepsilon_i.$$

Unfortunately, these methods quickly run into a “curse of dimensionality.” With  $r = 1$  predictor variable, it might take, say,  $K = 8$  parameters to get an adequate approximation to a regression

function  $m(\cdot)$ . With  $r = 2$  predictor variables, we could expect to need approximately  $K^2 = 8^2 = 64$  parameters in the linear model. Given a few hundred observations, that is doable. However, with  $r = 5$  predictor variables, we could expect to need about  $K^5 = 8^5 = 32,768$  parameters.

One way to avoid this problem is to fit *generalized additive models*, see Hastie et al. (2001). For example, with three predictor variables,  $x = (x_1, x_2, x_3)'$ , we might expect to need  $K^3 = 8^3 = 512$  terms to approximate  $m(\cdot)$ . To simplify the problem, we might assume that  $m(\cdot)$  follows a generalized additive model such as

$$m(x) = m_1(x_1) + m_{23}(x_2, x_3). \quad (4)$$

We might further approximate

$$m_1(x_1) = \sum_{k=1}^K \beta_{1k} \phi_k(x_1) \quad \text{and} \quad m_{23}(x_2, x_3) = \sum_{j=1}^K \sum_{k=1}^K \beta_{23jk} \phi_j(x_2) \phi_k(x_3).$$

If we need 8 terms to approximate  $m_1(\cdot)$  and 64 terms to approximate  $m_{23}(\cdot, \cdot)$ , the generalized additive model (4) involves fitting only 72 parameters rather than 512. With the 8 term approximations and 5 predictor variables, a generalized additive model that includes all of the possible  $m_{jk}(\cdot, \cdot)$ s involves only 640 terms, rather than the 32,768 required by a full implementation of a nonparametric regression.

**EXERCISE 15.11.** The function PSgam in DPpackage fits generalized additive models using B-splines. There is another predictor for the nitrogen oxides response  $y_i$  in the ethanol data besides the equivalence ratio  $x_i$ , namely the compression ratio  $z_i$ . Look at a scatterplot matrix of the three variables. Use the PSgam function to fit several models  $y_i = m(x_i, z_i) + \varepsilon_i$ : (a)  $m(x_i, z_i) = \beta_0 + m_1(x_i) + m_2(z_i)$  where  $m_1(\cdot)$  and  $m_2(\cdot)$  are both modeled using B-splines, (b)  $m(x_i, z_i) = \beta_0 + m_1(x_i)$  using equivalence ratio only as in all of the examples thus far, and (c)  $m(x_i, z_i) = \beta_0 + m_1(x_i) + \beta_2 z_i$ , nonlinear in  $x_i$  but linear in  $z_i$ . Compare the fits qualitatively and through LPML. Code is provided on our website in Chap15DPpackage.txt. Note that default plots obtained from PSgam remove the overall intercept  $\beta_0$  and force the functions  $m_i(\cdot)$  to integrate to 0 for identifiability, so the vertical scale of the plots will be different than the figures in this chapter for the ethanol data. See the comments on identifiability near the end of Section 3.

**EXERCISE 15.12.** The compression ratio  $z_i$  was observed at only 5 levels. What happens to the LPML when this variable is treated as categorical instead of continuous?

**EXERCISE 15.13.** Use the PSgam function to explore the “optimal” transformation of temperature in the O-ring data of Chapter 8 using a logistic link. Specifically, fit the model

$$\log \frac{\theta_i}{1 - \theta_i} = \beta_0 + m(T_i),$$

where  $T$  indicates temperature and  $m(T)$  is modeled using cubic B-splines with  $K = 20$ . Play with the prior on the penalty parameter  $\lambda \sim \text{Gamma}(\tau_{b1}, \tau_{b2})$ . How does the posterior trend change for priors that favor *large* values of  $\lambda$ ? Compare LPML for the best fitting B-spline model to LPML for simple regression models fitted using the logit, probit, and complementary log-log links.

### 15.3 Proportional Hazards Modeling

We now introduce alternatives to the step function model of Subsection 13.2.2 for the baseline hazard in Proportional Hazards models. Several authors (e.g., Royston, 2001; Hennerfeind et al.,

2006; Li, Hu, and Greene, 2009) have advocated modeling the log of the baseline hazard  $h_0(t)$  using flexible smooth functions such as those discussed in Section 2. Those methods are restricted to estimating functions over a finite interval which is a problem for estimating hazard functions. As discussed in Subsection 13.2.1, a hazard function must integrate to infinity. If the survival times have an upper bound, that means that the hazard function should approach infinity as time approaches the upper bound. None of the methods in Section 2 accommodate that requirement. As a practical matter, we just pick an upper bound that contains the observed data and admit that we have no idea what the hazard function looks like beyond that point.

BayesX (Belitz et al., 2009) is a flexible, easy to use, and free Windows-based program for fitting generalized additive mixed models with structured (e.g., spatially dependent) random effects, primarily written by Christiane Belitz, Andreas Brezger, Thomas Kneib, and Stefan Lang. It is available for download at <http://www.stat.uni-muenchen.de/~bayesx/bayesx.html>. BayesX provides computer functions for fitting proportional hazards models with the baseline hazard  $h_0(t)$  modeled as a piecewise constant like (13.2.4) or  $\log[h_0(t)]$  modeled using cubic B-splines. Generalizations include additive models, time-varying regression coefficients, time-dependent covariates, and exchangeable and spatially referenced random effects or frailty terms.

Without delving into great detail, that is, without discussing likelihoods and priors, we fit a PH model to the leukemia data of Chapter 12 using the B-spline option in BayesX. Write the model as

$$h(t|x, \beta) = h_0(t)e^{x\beta} = e^{\beta_0 + m_0(t)}e^{x\beta_1}, \quad (5)$$

where  $x = 0, 1$  denotes AG– or AG+. We also fit a non-PH time varying hazards ratio model

$$h(t|x, \beta) = e^{\beta_0 + m_1(t)}e^{xm_2(t)}. \quad (6)$$

Here  $m_0(t)$ ,  $m_1(t)$ , and  $m_2(t)$  are all modeled using cubic B-splines. The functions  $m_0(t)$  and  $m_1(t)$  integrate to 0 but  $m_2(t)$  has no such restriction. There are identifiability issues associated with these functions but they are completely unrelated to the identifiability issue associated with B-splines. These issues are discussed at the end of the section.

Posterior results for the proportional hazards model (5) include

node	mean	sd	2.5%	med	97.5%
$\beta_0$	-2.804	0.408	-3.641	-2.782	-1.993
$\beta_1$	-1.304	0.432	-2.138	-1.308	-0.445

Figure contains the posterior median along with 80% and 95% pointwise PIs for the zero-centered log-baseline hazard. Note the overall “bathtub” shape indicating elevated hazard at the beginning and the end of the study period. It appears to be just barely possible to fit a flat hazard (for an exponential model) between the 95% limits.

For the model (6), Figures and 15.16 give medians and intervals for  $m_1(t)$  and  $m_2(t)$ , respectively. For the simple two-group case, model (6) is equivalent to fitting two distinct hazard functions, one for each group. Figure gives estimated deviations from the “average” log-hazard for the AG– group. Adding the posterior median  $\tilde{\beta}_0 = -2.782$  to the posterior medians  $\tilde{m}_1(t)$  from Figure provides an estimate of the log-hazard for people with AG–. Figure 15.16 gives the deviations of AG+ log-hazards from the AG– log-hazards. The AG+ deviations tend to be negative, so log-hazards are lower for AG+s, reducing the AG+ hazards by a multiplicative effect relative to the AG– hazards, the amount of the reduction changes with time. With this model, we could easily have the AG+ differential log-hazards change from negative to positive over time, something that cannot happen in a proportional hazards model. Flat hazards fit easily between the 80% limits in Figures and 15.16, suggesting the plausibility of a two-group exponential model.

The DIC for the PH model (5) is 300.6 whereas the more complex model (6) gives 301.8. The simpler model is preferred, so the group effect can be described in a single hazard ratio estimated as

## Effect of time

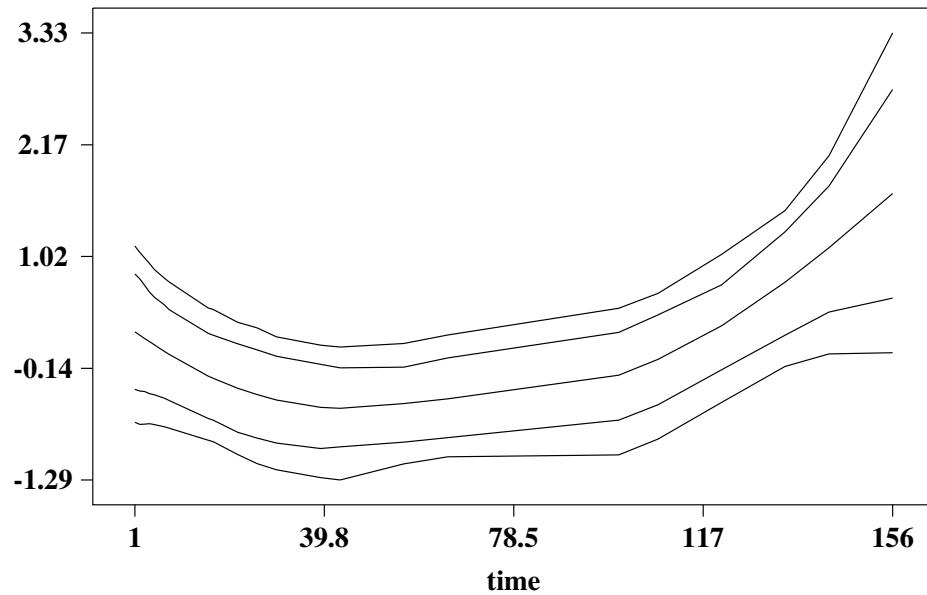


Figure 15.14: Zero-centered log-baseline hazard  $m_0(t)$  of model (5). Posterior median, 80%, and 95% intervals.

## Effect of time

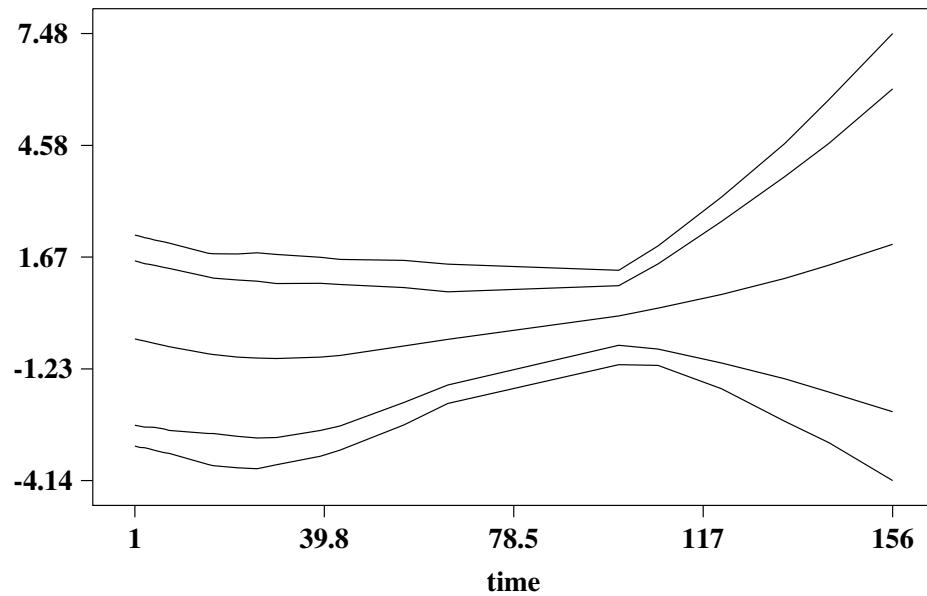


Figure 15.15: Zero-centered log-baseline hazard  $m_1(t)$  of model (6). Posterior median, 80%, and 95% intervals.

## Effect of group

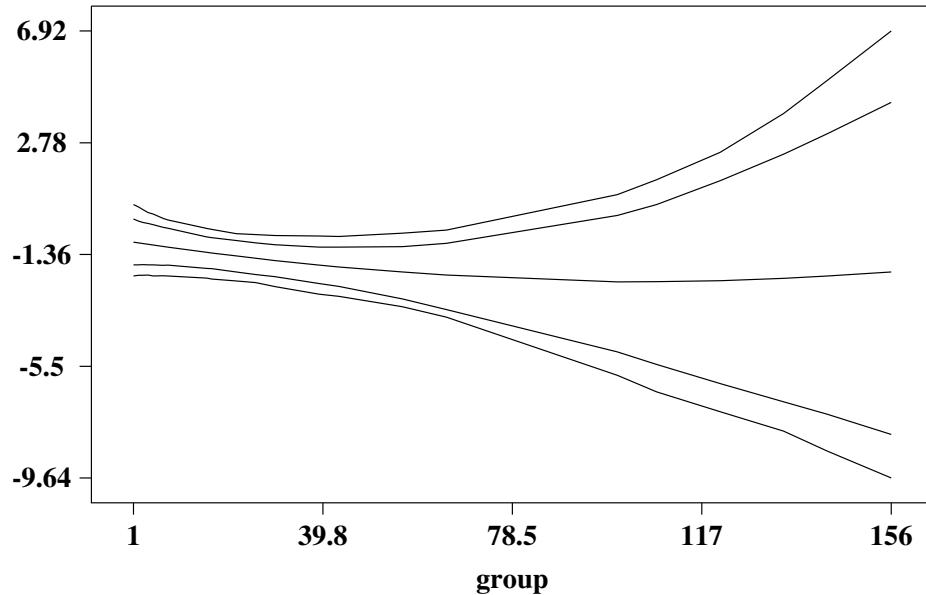


Figure 15.16: Varying group effect  $m_2(t)$  of model (6). Posterior median, 80%, and 95% intervals.

$e^{1.304} = 3.68$ . AG– increases the hazard of death by a factor of about 3.7. Recall that our estimate using the step function model (13.2.4) for  $h_0$  was 3.5.

The BayesX script for fitting these models is available on the book website.

As promised, we discuss the identifiability issues associated with model (6). Model (5) is simpler and easier. In model (6) with  $x = 0, 1$ , the overall log-hazard function is

$$m(t, x) = \beta_0 + m_1(t) + xm_2(t).$$

Here  $m(t, 0)$  is the log-hazard function for the AG– group and  $m(t, 1)$  is the log-hazard function for AG+ people. The function  $m(t, x)$  is identifiable because it uniquely determines the distributions of the survival times for both the AG– and AG+ groups. The function  $m_2$  is also identifiable because it is a function of identifiable quantities, i.e.,

$$m_2(t) = m(t, 1) - m(t, 0).$$

However,  $\beta_0$  and  $m_1$  are not identifiable because we only know

$$\beta_0 + m_1(t) = m(t, 0).$$

There are many choices for  $\beta_0$  and  $m_1(t)$  that will determine the same hazard function, and thus the same survival distribution for the AG– group. Imposing the artificial side condition  $\beta_0 \equiv \int m(t, 0) dt$  uniquely determines  $\beta_0$  and it follows that  $m_1(t) = m(t, 0) - \beta_0$  is both uniquely determined and integrates to 0.

Presumably, the goal of any estimation procedure for model (6) is to provide an estimate  $\hat{m}(t, x)$ . Obviously, this leads to  $\hat{m}_2(t) \equiv \hat{m}(t, 1) - \hat{m}(t, 0)$ . Estimates of  $\beta_0$  and  $m_1(t)$  depend on applying the artificial side condition to  $\hat{m}(t, 0)$ . The thought process is a bit like that for linear models. With  $Y = X\beta + e$ , the first order of business is to estimate  $E(Y)$ , which is the one thing that you can

obviously estimate (using  $Y$  if nothing else). Beyond that, parameters are identifiable or nonidentifiable depending on whether they are uniquely determined by  $E(Y)$ , see Christensen (2002, Sec. 2.1). Naturally, any estimate of  $E(Y)$  provides unique estimates of identifiable functions.

The methods in Sections 2 and 3 are closely related to functional data analysis, a subject to which Crainiceanu and Goldsmith (2009) provide a nice introduction.

This is the end of the last chapter. If you've completed this long, hard journey with us, all we can say is, "Good on ya, mate!" We hope you have learned as much from reading our book as we did writing it.

---

# Appendix A: Matrices and Vectors

---

This appendix gives the primary matrix facts used in the book. It is closely related to a similar appendix in Christensen (1996).

A matrix is a rectangular array of numbers. Such arrays have *rows* and *columns*. The numbers of rows and columns are referred to as the *dimensions* of a matrix. A matrix with, say, 5 rows and 3 columns is referred to as a  $5 \times 3$  matrix.

EXAMPLE A.0.1. Three matrices are given below along with their dimensions.

$$\begin{array}{c} \left[ \begin{array}{cc} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{array} \right], \quad \left[ \begin{array}{cc} 20 & 80 \\ 90 & 140 \end{array} \right], \quad \left[ \begin{array}{c} 6 \\ 180 \\ -3 \\ 0 \end{array} \right]. \\ 3 \times 2 \qquad \qquad \qquad 2 \times 2 \qquad \qquad \qquad 4 \times 1 \end{array}$$

Let  $r$  be an arbitrary positive integer. A matrix with  $r$  rows and  $r$  columns, i.e., an  $r \times r$  matrix, is called a *square matrix*. The second matrix in Example A.0.1 is square. A matrix with only one column, i.e., an  $r \times 1$  matrix, is a *vector*, sometimes called a *column vector*. The third matrix in Example A.0.1 is a vector. A  $1 \times r$  matrix is sometimes called a *row vector*.

An arbitrary matrix  $A$  is often written

$$A = [a_{ij}]$$

where  $a_{ij}$  denotes the element of  $A$  in the  $i$ th row and  $j$ th column. Two matrices are equal if they have the same dimensions and all of their elements (entries) are equal. Thus for  $r \times c$  matrices  $A = [a_{ij}]$  and  $B = [b_{ij}]$ ,  $A = B$  if and only if  $a_{ij} = b_{ij}$  for every  $i = 1, \dots, r$  and  $j = 1, \dots, c$ .

EXAMPLE A.0.2. Let

$$A = \begin{bmatrix} 20 & 80 \\ 90 & 140 \end{bmatrix} \text{ and } B = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}.$$

If  $B = A$ , then  $b_{11} = 20$ ,  $b_{12} = 80$ ,  $b_{21} = 90$ , and  $b_{22} = 140$ .

The *transpose* of a matrix  $A$ , denoted  $A'$ , changes the rows of  $A$  into columns of a new matrix  $A'$ . If  $A$  is an  $r \times c$  matrix, the transpose  $A'$  is a  $c \times r$  matrix. In particular, if we write  $A' = [\tilde{a}_{ij}]$ , then the element in row  $i$  and column  $j$  of  $A'$  is defined to be  $\tilde{a}_{ij} = a_{ji}$ .

EXAMPLE A.0.3.

$$\left[ \begin{array}{cc} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{array} \right]' = \left[ \begin{array}{ccc} 1 & 2 & 3 \\ 4 & 5 & 6 \end{array} \right]$$

and

$$\left[ \begin{array}{cc} 20 & 80 \\ 90 & 140 \end{array} \right]' = \left[ \begin{array}{cc} 20 & 90 \\ 80 & 140 \end{array} \right].$$

The transpose of a column vector is a row vector,

$$\begin{bmatrix} 6 \\ 180 \\ -3 \\ 0 \end{bmatrix}' = [6 \quad 180 \quad -3 \quad 0].$$

### A.1 Matrix Addition and Subtraction

Two matrices can be added (or subtracted) if they have the same dimensions, that is, if they have the same number of rows and columns. Addition and subtraction is performed elementwise.

EXAMPLE A.1.1.

$$\begin{aligned} \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix} + \begin{bmatrix} 2 & 8 \\ 4 & 10 \\ 6 & 12 \end{bmatrix} &= \begin{bmatrix} 1+2 & 4+8 \\ 2+4 & 5+10 \\ 3+6 & 6+12 \end{bmatrix} = \begin{bmatrix} 3 & 12 \\ 6 & 15 \\ 9 & 18 \end{bmatrix}. \\ \begin{bmatrix} 20 & 80 \\ 90 & 140 \end{bmatrix} - \begin{bmatrix} -15 & -75 \\ 80 & 130 \end{bmatrix} &= \begin{bmatrix} 35 & 155 \\ 10 & 10 \end{bmatrix}. \end{aligned}$$

In general, if  $A$  and  $B$  are  $r \times c$  matrices with  $A = [a_{ij}]$  and  $B = [b_{ij}]$ , then

$$A + B = [a_{ij} + b_{ij}] \text{ and } A - B = [a_{ij} - b_{ij}].$$

### A.2 Scalar Multiplication

Any matrix can be multiplied by a scalar. Multiplication by a scalar (a *real number*) is elementwise.

EXAMPLE A.2.1. Scalar multiplication gives

$$\begin{aligned} \frac{1}{10} \begin{bmatrix} 20 & 80 \\ 90 & 140 \end{bmatrix} &= \begin{bmatrix} 20/10 & 80/10 \\ 90/10 & 140/10 \end{bmatrix} = \begin{bmatrix} 2 & 8 \\ 9 & 14 \end{bmatrix}. \\ 2[6 \quad 180 \quad -3 \quad 0] &= [12 \quad 360 \quad -6 \quad 0]. \end{aligned}$$

In general, if  $\lambda$  is any number and  $A = [a_{ij}]$ , then

$$\lambda A = [\lambda a_{ij}].$$

### A.3 Matrix Multiplication

Two matrices can be multiplied together if the number of columns in the first matrix is the same as the number of rows in the second matrix. In the process of multiplication, the rows of the first matrix are matched up with the columns of the second matrix.

EXAMPLE A.3.1.

$$\begin{aligned} \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix} \begin{bmatrix} 20 & 80 \\ 90 & 140 \end{bmatrix} &= \begin{bmatrix} (1)(20) + (4)(90) & (1)(80) + (4)(140) \\ (2)(20) + (5)(90) & (2)(80) + (5)(140) \\ (3)(20) + (6)(90) & (3)(80) + (6)(140) \end{bmatrix} \\ &= \begin{bmatrix} 380 & 640 \\ 490 & 860 \\ 600 & 1080 \end{bmatrix}. \end{aligned}$$

The entry in the first row and column of the product matrix,  $(1)(20) + (4)(90)$ , matches the elements in the first row of the first matrix,  $(1\ 4)$ , with the elements in the first column of the second matrix,  $\begin{pmatrix} 20 \\ 90 \end{pmatrix}$ . The 1 in  $(1\ 4)$  is matched up with the 20 in  $\begin{pmatrix} 20 \\ 90 \end{pmatrix}$  and these numbers are multiplied.

Similarly, the 4 in  $(1\ 4)$  is matched up with the 90 in  $\begin{pmatrix} 20 \\ 90 \end{pmatrix}$  and the numbers are multiplied. Finally, the two products are added to obtain the entry  $(1)(20) + (4)(90)$ . Similarly, the entry in the third row, second column of the product,  $(3)(80) + (6)(140)$ , matches the elements in the third row of the first matrix,  $(3\ 6)$ , with the elements in the second column of the second matrix,  $\begin{pmatrix} 80 \\ 140 \end{pmatrix}$ . After multiplying and adding we get the entry  $(3)(80) + (6)(140)$ . To carry out this matching, the number of columns in the first matrix must equal the number of rows in the second matrix. The matrix product has the same number of rows as the first matrix and the same number of columns as the second because each row of the first matrix can be matched with each column of the second.

**EXAMPLE A.3.2.** We illustrate another matrix multiplication commonly performed in statistics, multiplying a matrix on its left by the transpose of that matrix, i.e., computing  $A'A$ .

$$\begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}' \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix} \\ = \begin{bmatrix} 1+4+9 & 4+10+18 \\ 4+10+18 & 16+25+36 \end{bmatrix} \\ = \begin{bmatrix} 14 & 32 \\ 32 & 77 \end{bmatrix}.$$

Notice that in matrix multiplication the roles of the first matrix and the second matrix are *not* interchangeable. In particular, if we reverse the order of the matrices in Example A.3.1, the matrix product

$$\begin{bmatrix} 20 & 80 \\ 90 & 140 \end{bmatrix} \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$$

is undefined because the first matrix has two columns while the second matrix has three rows. Even when the matrix products are defined for both  $AB$  and  $BA$ , the results of the multiplication typically differ. If  $A$  is  $r \times s$  and  $B$  is  $s \times r$ , then  $AB$  is an  $r \times r$  matrix and  $BA$  is an  $s \times s$  matrix. When  $r \neq s$ , clearly  $AB \neq BA$ , but even when  $r = s$  we still cannot expect  $AB$  to equal  $BA$ .

**EXAMPLE A.3.3.** Consider two square matrices, say,

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \quad B = \begin{bmatrix} 0 & 2 \\ 1 & 2 \end{bmatrix}.$$

Multiplication gives

$$AB = \begin{bmatrix} 2 & 6 \\ 4 & 14 \end{bmatrix}$$

and

$$BA = \begin{bmatrix} 6 & 8 \\ 7 & 10 \end{bmatrix},$$

so  $AB \neq BA$ .

In general if  $A = [a_{ij}]$  is an  $r \times s$  matrix and  $B = [b_{ij}]$  is an  $s \times c$  matrix, then

$$AB = [d_{ij}]$$

is the  $r \times c$  matrix with

$$d_{ij} = \sum_{\ell=1}^s a_{i\ell} b_{\ell j}.$$

A useful result is that the transpose of the product  $AB$  is the product, in reverse order, of the transposed matrices, i.e.,  $(AB)' = B'A'$ .

**EXAMPLE A.3.4.** As seen in Example A.3.1,

$$AB \equiv \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix} \begin{bmatrix} 20 & 80 \\ 90 & 140 \end{bmatrix} = \begin{bmatrix} 380 & 640 \\ 490 & 860 \\ 600 & 1080 \end{bmatrix} \equiv C.$$

The transpose of this matrix is

$$C' = \begin{bmatrix} 380 & 490 & 600 \\ 640 & 860 & 1080 \end{bmatrix} = \begin{bmatrix} 20 & 90 \\ 80 & 140 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} = B'A'.$$

Let  $a = (a_1, \dots, a_n)'$  be a vector. A very useful property of vectors is that

$$a'a = \sum_{i=1}^n a_i^2 \geq 0.$$

#### A.4 Special Matrices

If  $A = A'$ , then  $A$  is said to be *symmetric*. If  $A = [a_{ij}]$  and  $A = A'$ , then  $a_{ij} = a_{ji}$ . The entry in row  $i$  and column  $j$  is the same as the entry in row  $j$  and column  $i$ . Only square matrices can be symmetric.

**EXAMPLE A.4.1.** The matrix

$$A = \begin{bmatrix} 4 & 3 & 1 \\ 3 & 2 & 6 \\ 1 & 6 & 5 \end{bmatrix}$$

has  $A = A'$ .  $A$  is symmetric about the diagonal that runs from the upper left to the lower right.

For any  $r \times c$  matrix  $A$ , the product  $A'A$  is always symmetric. This was illustrated in Example A.3.2. More generally, write  $A = [a_{ij}]$ ,  $A' = [\tilde{a}_{ij}]$  with  $\tilde{a}_{ij} = a_{ji}$ , and

$$A'A = [d_{ij}] = \left[ \sum_{\ell=1}^c \tilde{a}_{i\ell} a_{\ell j} \right].$$

Note that

$$d_{ij} = \sum_{\ell=1}^c \tilde{a}_{i\ell} a_{\ell j} = \sum_{\ell=1}^c a_{\ell i} a_{\ell j} = \sum_{\ell=1}^c \tilde{a}_{j\ell} a_{\ell i} = d_{ji}$$

so the matrix is symmetric.

*Diagonal matrices* are square matrices with all off diagonal elements equal to zero.

**EXAMPLE A.4.2.** The matrices

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}, \quad \begin{bmatrix} 20 & 0 \\ 0 & -3 \end{bmatrix}, \text{ and } \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

are diagonal.

In general, a diagonal matrix is a square matrix  $A = [a_{ij}]$  with  $a_{ij} = 0$  for  $i \neq j$ . Obviously, diagonally matrices are symmetric.

An *identity matrix* is a diagonal matrix with all 1s along the diagonal, i.e.,  $a_{ii} = 1$  for all  $i$ . The third matrix in Example A.4.2 above is a  $3 \times 3$  identity matrix. The identity matrix gets its name because any matrix multiplied by an identity matrix remains unchanged.

EXAMPLE A.4.3.

$$\begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}.$$

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix} = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}.$$

An  $r \times r$  identity matrix is denoted  $I_r$  with the subscript deleted if the dimension is clear.

A *zero matrix* is a matrix that consists entirely of zeros. Obviously, the product of any matrix multiplied by a zero matrix is zero.

EXAMPLE A.4.4.

$$\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

Often a zero matrix is denoted by 0 where the dimension of the matrix, and the fact that it is a matrix rather than a scalar, must be inferred from the context.

## A.5 Linear Dependence and Rank

Consider the matrix

$$A = \begin{bmatrix} 1 & 2 & 5 & 1 \\ 2 & 2 & 10 & 6 \\ 3 & 4 & 15 & 1 \end{bmatrix}.$$

Note that each column of  $A$  can be viewed as a vector. The *column space* of  $A$ , denoted  $C(A)$ , is the collection of all vectors that can be written as a *linear combination of the columns of  $A$* . In other words,  $C(A)$  is the set of all vectors that can be written as

$$\lambda_1 \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} + \lambda_2 \begin{bmatrix} 2 \\ 2 \\ 4 \end{bmatrix} + \lambda_3 \begin{bmatrix} 5 \\ 10 \\ 15 \end{bmatrix} + \lambda_4 \begin{bmatrix} 1 \\ 6 \\ 1 \end{bmatrix} = A \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \end{bmatrix} = A\lambda$$

for some vector  $\lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4)'$ .

The columns of any matrix  $A$  are *linearly dependent* if they contain redundant information. Specifically, let  $x$  be some vector in  $C(A)$ . The columns of  $A$  are linearly dependent if we can find two distinct vectors  $\lambda$  and  $\gamma$  such that  $x = A\lambda$  and  $x = A\gamma$ . Thus two distinct linear combinations of the columns of  $A$  give rise to the same vector  $x$ . Note that  $\lambda \neq \gamma$  because  $\lambda$  and  $\gamma$  are distinct. Note also that, using a distributive property of matrix multiplication,  $A(\lambda - \gamma) = A\lambda - A\gamma = 0$ , where  $\lambda - \gamma \neq 0$ . This condition is frequently used as an alternative definition for linear dependence, i.e., the columns of  $A$  are linearly dependent if there exists a vector  $\delta \neq 0$  such that  $A\delta = 0$ . If the columns of  $A$  are not linearly dependent, they are *linearly independent*.

EXAMPLE A.5.1. Observe that the example matrix  $A$  given at the beginning of the section has

$$\begin{bmatrix} 1 & 2 & 5 & 1 \\ 2 & 2 & 10 & 6 \\ 3 & 4 & 15 & 1 \end{bmatrix} \begin{bmatrix} 5 \\ 0 \\ -1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix},$$

so the columns of  $A$  are linearly dependent.

The *rank* of  $A$  is the smallest number of columns of  $A$  that can generate  $C(A)$ . It is also the maximum number of linearly independent columns in  $A$ .

EXAMPLE A.5.2. The matrix

$$A = \begin{bmatrix} 1 & 2 & 5 & 1 \\ 2 & 2 & 10 & 6 \\ 3 & 4 & 15 & 1 \end{bmatrix}$$

has rank 3 because the columns

$$\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 2 \\ 2 \\ 4 \end{bmatrix}, \begin{bmatrix} 1 \\ 6 \\ 1 \end{bmatrix}$$

generate  $C(A)$ . We saw in Example A.5.1 that the column  $(5, 10, 15)'$  was redundant. None of the other three columns are redundant; they are linearly independent. In other words, the only way to get

$$\begin{bmatrix} 1 & 2 & 1 \\ 2 & 2 & 6 \\ 3 & 4 & 1 \end{bmatrix} \delta = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

is to take  $\delta = (0, 0, 0)'$ .

## A.6 Inverse Matrices

The *inverse* of a square matrix  $A$  is the matrix  $A^{-1}$  such that

$$AA^{-1} = A^{-1}A = I.$$

The inverse of  $A$  exists only if the columns of  $A$  are linearly independent. Typically, it is difficult to find inverses without the aid of a computer. For a  $2 \times 2$  matrix

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix},$$

the inverse is given by

$$A^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}. \quad (1)$$

To confirm that this is correct, multiply  $AA^{-1}$  to see that it gives the identity matrix. Moderately complicated formulae exist for computing the inverse of  $3 \times 3$  matrices. Inverses of larger matrices become very difficult to compute by hand. Of course computers are ideally suited for finding such things.

One use for inverse matrices is in solving systems of equations.

EXAMPLE A.6.1. Consider the system of equations

$$\begin{aligned} 2x + 4y &= 20 \\ 3x + 4y &= 10. \end{aligned}$$

We can write this in matrix form as

$$\begin{bmatrix} 2 & 4 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 20 \\ 10 \end{bmatrix}.$$

Multiplying on the left by the inverse of the coefficient matrix gives

$$\begin{bmatrix} 2 & 4 \\ 3 & 4 \end{bmatrix}^{-1} \begin{bmatrix} 2 & 4 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2 & 4 \\ 3 & 4 \end{bmatrix}^{-1} \begin{bmatrix} 20 \\ 10 \end{bmatrix}.$$

Using the definition of the inverse on the left-hand side of the equality and the formula in (A.6.1) on the right-hand side gives

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -1 & 1 \\ 3/4 & -1/2 \end{bmatrix} \begin{bmatrix} 20 \\ 10 \end{bmatrix}$$

or

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -10 \\ 10 \end{bmatrix}.$$

Thus  $(x, y) = (-10, 10)$  is the solution for the two equations, i.e.,  $2(-10) + 4(10) = 20$  and  $3(-10) + 4(10) = 10$ .

More generally a system of equations, say,

$$\begin{aligned} a_{11}y_1 + a_{12}y_2 + a_{13}y_3 &= c_1 \\ a_{21}y_1 + a_{22}y_2 + a_{23}y_3 &= c_2 \\ a_{31}y_1 + a_{32}y_2 + a_{33}y_3 &= c_3 \end{aligned}$$

in which the  $a_{ij}$ s and  $c_i$ s are known and the  $y_i$ s are variables, can be written in matrix form as

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix}$$

or

$$AY = C.$$

To find  $Y$  simply observe that  $AY = C$  implies  $A^{-1}AY = A^{-1}C$  and  $Y = A^{-1}C$ . Of course this argument assumes that  $A^{-1}$  exists, which is not always the case. Moreover, the procedure obviously extends to larger sets of equations.

On a computer, there are better ways of finding solutions to systems of equations than finding the inverse of a matrix. In fact, inverses are often found by solving systems of equations. For example, in a  $3 \times 3$  case the first column of  $A^{-1}$  can be found as the solution to

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$

For a special type of square matrix, called an *orthogonal matrix*, the transpose is also the inverse. In other words, a square matrix  $P$  is an orthogonal matrix if

$$P'P = I = PP'.$$

To establish that  $P$  is orthogonal, it is enough to show either that  $P'P = I$  or that  $PP' = I$ . Orthogonal matrices are particularly useful in discussions of eigenvalues and eigenvectors.

### A.7 A List of Useful Properties

The following proposition summarizes many of the key properties of matrices and the operations performed on them.

**Proposition A.7.1.** Let  $A$ ,  $B$ , and  $C$  be matrices of appropriate dimensions and let  $\lambda$  be a scalar.

$$\begin{aligned} A + B &= B + A \\ (A + B) + C &= A + (B + C) \\ (AB)C &= A(BC) \\ C(A + B) &= CA + CB \\ \lambda(A + B) &= \lambda A + \lambda B \\ (A')' &= A \\ (A + B)' &= A' + B' \\ (AB)' &= B'A' \\ (A^{-1})^{-1} &= A \\ (A')^{-1} &= (A^{-1})' \\ (AB)^{-1} &= B^{-1}A^{-1}. \end{aligned}$$

The last equality only holds when  $A$  and  $B$  both have inverses. The second to the last property implies that the inverse of a symmetric matrix is symmetric because then  $A^{-1} = (A')^{-1} = (A^{-1})'$ . This is a very important property.

### A.8 Eigenvalues and Eigenvectors

Let  $A$  be a square matrix. A scalar  $\phi$  is an eigenvalue of  $A$  and  $x \neq 0$  is an eigenvector for  $A$  corresponding to  $\phi$  if

$$Ax = \phi x.$$

EXAMPLE A.8.1. Consider the matrix

$$A = \begin{bmatrix} 3 & 1 & -1 \\ 1 & 3 & -1 \\ -1 & -1 & 5 \end{bmatrix}.$$

The value 3 is an eigenvalue and any nonzero multiple of the vector  $(1, 1, 1)'$  is a corresponding eigenvector. For example,

$$\begin{bmatrix} 3 & 1 & -1 \\ 1 & 3 & -1 \\ -1 & -1 & 5 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \\ 3 \end{bmatrix} = 3 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

Similarly, if we consider a multiple, say,  $4(1, 1, 1)'$ ,

$$\begin{bmatrix} 3 & 1 & -1 \\ 1 & 3 & -1 \\ -1 & -1 & 5 \end{bmatrix} \begin{bmatrix} 4 \\ 4 \\ 4 \end{bmatrix} = \begin{bmatrix} 12 \\ 12 \\ 12 \end{bmatrix} = 3 \begin{bmatrix} 4 \\ 4 \\ 4 \end{bmatrix}.$$

The value 2 is also an eigenvalue with eigenvectors that are nonzero multiples of  $(1, -1, 0)'$ .

$$\begin{bmatrix} 3 & 1 & -1 \\ 1 & 3 & -1 \\ -1 & -1 & 5 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} = \begin{bmatrix} 2 \\ -2 \\ 0 \end{bmatrix} = 2 \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}.$$

Finally, 6 is an eigenvalue with eigenvectors that are nonzero multiples of  $(1, 1, -2)'$ .

**Proposition A.8.2.** Let  $A$  be a symmetric matrix. Then for a diagonal matrix  $D(\phi_i)$  consisting of eigenvalues there exists an orthogonal matrix  $P$  whose columns are corresponding eigenvectors such that

$$A = PD(\phi_i)P'.$$

EXAMPLE A.8.3. Consider again the matrix

$$A = \begin{bmatrix} 3 & 1 & -1 \\ 1 & 3 & -1 \\ -1 & -1 & 5 \end{bmatrix}.$$

In writing  $A = PD(\phi_i)P'$ , the diagonal matrix is

$$D(\phi_i) = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 6 \end{bmatrix}.$$

The orthogonal matrix is

$$P = \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & 0 & \frac{-2}{\sqrt{6}} \end{bmatrix}.$$

We leave it to the reader to verify that  $PD(\phi_i)P' = A$  and that  $P'P = I$ .

Note that the columns of  $P$  are multiples of the vectors identified as eigenvectors in Example A.8.1; hence, the columns of  $P$  are also eigenvectors. The multiples of the eigenvectors were chosen so that  $PP' = I$  and  $P'P = I$ . Moreover, the first column of  $P$  is an eigenvector corresponding to 3, which is the first eigenvalue listed in  $D(\phi_i)$ . Similarly, the second column of  $P$  is an eigenvector corresponding to 2 and the third column corresponds to the third listed eigenvalue, 6.

For a  $3 \times 3$  matrix  $A$  that has three *distinct* eigenvalues, any matrix  $P$  with eigenvectors for columns would have  $P'P$  a diagonal matrix, but the multiples of the eigenvectors must be chosen so that the diagonal entries of  $P'P$  are all 1.

EXAMPLE A.8.4. Consider the matrix

$$B = \begin{bmatrix} 5 & -1 & -1 \\ -1 & 5 & -1 \\ -1 & -1 & 5 \end{bmatrix}.$$

This matrix is closely related to the matrix in Example A.8.1. The matrix  $B$  has 3 as an eigenvalue with corresponding eigenvectors that are multiples of  $(1, 1, 1)'$ , just like the matrix  $A$ . Once again 6 is an eigenvalue with corresponding eigenvector  $(1, 1, -2)'$  and once again  $(1, -1, 0)'$  is an eigenvector, but now, unlike  $A$ ,  $(1, -1, 0)$  also corresponds to the eigenvalue 6. We leave it to the reader to verify these facts. The point is that in this matrix, 6 is an eigenvalue that has two linearly independent eigenvectors. In such cases, any nonzero linear combination of the two eigenvectors is also an eigenvector. For example, it is easy to see that

$$3 \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} + 2 \begin{bmatrix} 1 \\ 1 \\ -2 \end{bmatrix} = \begin{bmatrix} 5 \\ -1 \\ -4 \end{bmatrix}$$

is an eigenvector corresponding to the eigenvalue 6.

To write  $B = PD(\phi)P'$  as in Proposition A.8.2,  $D(\phi)$  has 3, 6, and 6 down the diagonal and one choice of  $P$  is that given in Example A.8.3. However, because one of the eigenvalues occurs more than once in the diagonal matrix, there are many choices for  $P$ .

Generally, if we need eigenvalues or eigenvectors we get a computer to find them for us.

Two frequently used functions of a square matrix are the determinant and the trace.

#### Definition A.8.5.

- a) The determinant of a square matrix is the product of the eigenvalues of the matrix.
- b) The trace of a square matrix is the sum of the eigenvalues of the matrix.

In fact, one can show that the trace of a square matrix also equals the sum of the diagonal elements of that matrix.

#### A.9 Properties of Determinants

The determinant of a square matrix  $A$  is denoted  $\det(A)$ . If  $A$  and  $B$  are square and  $V = AB$ , then  $\det(V) = \det(AB) = \det(A)\det(B)$ . If  $A$  is nonsingular,  $I = AA^{-1}$ , so  $1 = \det(I) = \det(A)\det(A^{-1})$  and  $\det(A^{-1}) = 1/\det(A)$ .

#### A.10 Calculus and Taylor's Theorem

Consider transforming an  $n$  dimensional vector  $y$  into a  $t$  dimensional vector  $z = G(y)$ . Write  $G(y)$  as a vector

$$G(y) = \begin{bmatrix} g_1(y) \\ \vdots \\ g_t(y) \end{bmatrix}.$$

The derivative of  $G$  is  $dG(y)$  which is defined to be the  $n \times t$  matrix of partial derivatives  $[\partial g_j(y)/\partial y_i]$ . In particular, if  $G$  is a linear transformation, say  $G(y) = A'y$  for some fixed  $n \times t$  matrix  $A$ , it is not difficult to see that  $dG(y) = A$ .

Let  $\theta$  be a vector  $\theta = (\theta_1, \dots, \theta_r)'$ . Consider a function  $\ell(\theta)$  that maps  $\mathbf{R}^r$  into  $\mathbf{R}$ . Let  $\hat{\theta}$  be a fixed  $r$  vector. Let  $\dot{\ell}(\theta)$  be the  $r$  dimensional vector of partial derivatives of  $\ell(\theta)$  with respect to the  $\theta_i$ s. Let  $\ddot{\ell}(\theta)$  be the symmetric matrix of second order partial derivatives  $[\partial^2 \ell(\theta)/\partial \theta_i \partial \theta_j]$ . Taylor's Theorem tells us that for  $\theta$ s in a neighborhood of any point  $\hat{\theta}$  we can approximate  $\ell(\theta)$ . A first order Taylor's approximation is

$$\ell(\theta) \doteq \ell(\hat{\theta}) + \dot{\ell}(\hat{\theta})'(\theta - \hat{\theta}).$$

A second order Taylor's approximation is

$$\ell(\theta) \doteq \ell(\hat{\theta}) + \dot{\ell}(\hat{\theta})'(\theta - \hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})'\ddot{\ell}(\hat{\theta})(\theta - \hat{\theta}).$$

#### A.11 Partitioned Matrices

Suppose we have two matrices with the same number of rows,

$$A_1 = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 0 \end{bmatrix}; \quad A_2 = \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix}.$$

We can create a new matrix, a *partitioned matrix*,

$$A \equiv [A_1, A_2] = \begin{bmatrix} 1 & 4 & 1 \\ 2 & 5 & 0 \\ 3 & 0 & 2 \end{bmatrix}.$$

Similarly, if we have two matrices with the same number of columns,

$$B_1 = \begin{bmatrix} 1 & 2 \\ 1 & 3 \end{bmatrix}, \quad B_2 = \begin{bmatrix} 2 & 4 \end{bmatrix},$$

we can create

$$B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 2 & 4 \end{bmatrix}.$$

Most interestingly, if the numbers of columns and rows match up appropriately, we can multiply

$$AB = [A_1, A_2] \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} = A_1 B_1 + A_2 B_2.$$

To see this in our example, observe that

$$\begin{aligned} AB &= \begin{bmatrix} 1 & 4 & 1 \\ 2 & 5 & 0 \\ 3 & 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 2 & 4 \end{bmatrix} = \begin{bmatrix} 7 & 18 \\ 7 & 19 \\ 7 & 14 \end{bmatrix} \\ A_1 B_1 &= \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 0 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} 5 & 14 \\ 7 & 19 \\ 3 & 6 \end{bmatrix} \\ A_2 B_2 &= \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix} [2, 4] = \begin{bmatrix} 2 & 4 \\ 0 & 0 \\ 4 & 8 \end{bmatrix}. \end{aligned}$$

Also observe that

$$\begin{aligned} A' &= \begin{bmatrix} A'_1 \\ A'_2 \end{bmatrix}, \\ A'A &= \begin{bmatrix} A'_1 \\ A'_2 \end{bmatrix} [A_1, A_2] = \begin{bmatrix} A'_1 A_1 & A'_1 A_2 \\ A'_2 A_1 & A'_2 A_2 \end{bmatrix}, \\ AA' &= [A_1, A_2] \begin{bmatrix} A'_1 \\ A'_2 \end{bmatrix} = A_1 A'_1 + A_2 A'_2. \end{aligned}$$



---

# Appendix B: Probability

---

We assume that the reader has prior exposure to calculus-based probability and multivariable calculus. This appendix gives the primary probabilistic facts used in the book. *It is convenient for us to refer to both probability mass functions and probability density functions as density functions. If we need to make a distinction, a probability mass function will be referred to as a discrete density.*

## B.1 Univariate Probability

Let  $\theta$  be a random variable and  $u$  a number or place holder variable. The *cumulative distribution function (cdf)* of  $\theta$  is

$$F_\theta(u) \equiv \Pr[\theta \leq u].$$

If  $F_\theta(u)$  is the cdf of a discrete random variable, we define a *discrete density*

$$p_\theta(u) \equiv \Pr[\theta = u].$$

If  $F_\theta(u)$  admits a derivative, the random variable is continuous and we can define a *density* function

$$p_\theta(u) \equiv \frac{d}{du} F_\theta(u).$$

We give results as if for continuous densities using integration. For discrete densities, integrals are replaced by sums.

The cdf can be recovered from the density as

$$F_\theta(u) = \int_{-\infty}^u p_\theta(w) dw.$$

The expected value of  $\theta$  is defined as

$$\mathbb{E}[\theta] \equiv \int_{-\infty}^{\infty} up_\theta(u) du.$$

More generally, for a function  $g(\cdot)$  mapping  $\theta$  into  $\mathbf{R}$ , the expected value of  $g(\theta)$  is defined as

$$\mathbb{E}[g(\theta)] \equiv \int_{-\infty}^{\infty} g(u)p_\theta(u) du.$$

In particular, if we let  $\theta_0$  be  $\mathbb{E}[\theta]$ , the variance of  $\theta$  is defined to be

$$\text{Var}(\theta) \equiv \mathbb{E}[(\theta - \theta_0)^2] = \int_{-\infty}^{\infty} (u - \theta_0)^2 p_\theta(u) du.$$

In this appendix we write density functions  $p_\theta(u)$  with the subscript  $\theta$  indicating the random variable and a placeholder variable  $u$ . In most of the book, it will be more convenient to write the density as  $p(\theta)$ . For example, a commonly used continuous distribution is the Gamma distribution with parameters  $a$  and  $b$ . To indicate that  $\theta$  has such a Gamma distribution, we write

$$\theta \sim \text{Gamma}(a, b).$$

The density is written

$$p(\theta) = \frac{b^a}{\Gamma(a)} \theta^{a-1} \exp(-b\theta) I_{(0,\infty)}(\theta).$$

Here  $I_{(0,\infty)}(\theta)$  is the indicator function, that is, for a set  $A$

$$I_A(\theta) = \begin{cases} 1 & \text{if } \theta \in A \\ 0 & \text{otherwise} \end{cases}.$$

## B.2 Multivariate Probability

We now give a brief introduction to vectors and matrices that are made up of random variables and the probability calculations associated with them.

Let  $\theta_1, \dots, \theta_r$  be random variables. Define a random vector  $\theta = (\theta_1, \dots, \theta_r)'$ . Let  $u = (u_1, \dots, u_r)'$  be a vector of numbers or placeholder variables. The *joint cumulative distribution function* (cdf) of  $(\theta_1, \dots, \theta_r)'$  is

$$F_\theta(u) \equiv F_\theta(u_1, \dots, u_r) \equiv \Pr[\theta_1 \leq u_1, \dots, \theta_r \leq u_r].$$

If  $F_\theta(u_1, \dots, u_r)$  is the cdf of a discrete random variable, we can define a (*joint*) *discrete density*

$$p_\theta(u) = p_\theta(u_1, \dots, u_r) \equiv \Pr[\theta_1 = u_1, \dots, \theta_r = u_r].$$

If  $F_\theta(u_1, \dots, u_r)$  admits the  $r$ th order mixed partial derivative, the random vector is continuous and we can define a (*joint*) *density* function

$$p_\theta(u) = p_\theta(u_1, \dots, u_r) \equiv \frac{\partial^n}{\partial u_1 \cdots \partial u_r} F_\theta(u_1, \dots, u_r).$$

The *support* of the distribution is  $\{u | p_\theta(u) > 0\}$ , so the support has probability 1 and the complement of the support has probability 0.

All of the results presented work essentially the same way for combinations of discrete and continuous random variables, but to establish such results we would need to use measure theory, something we would rather avoid.

The cdf can be recovered from the density as

$$F_\theta(u_1, \dots, u_r) = \int_{-\infty}^{u_1} \cdots \int_{-\infty}^{u_r} p_\theta(w_1, \dots, w_r) dw_1 \cdots dw_r.$$

For a function  $g(\cdot)$  of  $(\theta_1, \dots, \theta_r)'$  into  $\mathbf{R}$ , the expected value is defined as

$$\mathbb{E}[g(\theta_1, \dots, \theta_r)] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(u_1, \dots, u_r) p_\theta(u_1, \dots, u_r) du_1 \cdots du_r.$$

Expectations have a linearity property. For  $a$  and  $b$  real numbers and real valued functions  $g_1$  and  $g_2$ ,

$$\mathbb{E}[ag_1(\theta) + bg_2(\theta)] = a\mathbb{E}[g_1(\theta)] + b\mathbb{E}[g_2(\theta)].$$

Let  $y_1, y_2, y_3$ , and  $y_4$  be random variables. Construct a  $4 \times 1$  random vector, say,

$$y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}.$$

The expected value of the random vector is the vector of expected values for the random variables. Write  $E(y_i) = \mu_i$ , then

$$E(y) \equiv \begin{bmatrix} E(y_1) \\ E(y_2) \\ E(y_3) \\ E(y_4) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{bmatrix} \equiv \mu.$$

Thus, the expectation of random vectors is performed elementwise.

In fact, the expected value of any random matrix (a matrix consisting of random variables) is the matrix of expected values of the elements in the random matrix. Thus, for  $w_{ij}$ ,  $i = 1, 2, 3$ ,  $j = 1, 2$  a collection of random variables, write

$$W = \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \\ w_{31} & w_{33} \end{bmatrix},$$

then

$$E(W) \equiv \begin{bmatrix} E(w_{11}) & E(w_{12}) \\ E(w_{21}) & E(w_{22}) \\ E(w_{31}) & E(w_{33}) \end{bmatrix}.$$

We need an idea for random vectors corresponding to the variance of a random variable. This is the *covariance matrix*, also known as the dispersion matrix, the variance matrix, or the variance-covariance matrix. The covariance matrix is a matrix of all the variances and covariances associated with  $y$ . Write

$$\text{Var}(y_i) = E(y_i - \mu_i)^2 \equiv \sigma_{ii}$$

and define the covariance between  $y_i$  and  $y_j$  as

$$\text{Cov}(y_i, y_j) = E[(y_i - \mu_i)(y_j - \mu_j)] \equiv \sigma_{ij}.$$

Two subscripts are used on  $\sigma_{ii}$  to indicate that it is the variance of  $y_i$  *rather than* writing  $\text{Var}(y_i) = \sigma_i^2$ . The covariance matrix of our  $4 \times 1$  vector  $y$  is

$$\text{Cov}(y) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} \end{bmatrix}.$$

When  $y$  is  $4 \times 1$ , the covariance matrix is  $4 \times 4$ . If  $y$  were  $30 \times 1$ ,  $\text{Cov}(y)$  would be  $30 \times 30$ . The covariance matrix is always symmetric because  $\sigma_{ij} = \sigma_{ji}$  for any  $i, j$ . The variances of the individual random variables lie on the diagonal going from the top left to the bottom right. Covariances lie off the diagonal.

In general, if  $y$  is an  $n \times 1$  random vector and  $E(y) = \mu$ , then  $\text{Cov}(y) = E[(y - \mu)(y - \mu)']$ . In other words,  $\text{Cov}(y)$  is the expected value of the random matrix  $(y - \mu)(y - \mu)'$ .

Using the linearity property of expectations, it is not difficult to see that if  $z$  is a random  $r$  vector with  $E(z) = \mu_z$  and  $\text{Cov}(z) = V_z$ , and if we transform  $z$  into  $y$  through  $y = Az + b$  where  $A$  is a fixed  $n \times r$  nonsingular matrix and  $b$  is a fixed  $n$  vector, then

$$E(y) = E(Az + b) = AE(z) + b = A\mu_z + b$$

and

$$\begin{aligned} \text{Cov}(y) &= E[(Az + b) - (A\mu_z + b)][(Az + b) - (A\mu_z + b)]' \\ &= E[A(z - \mu_z)][A(z - \mu_z)']' \\ &= E[A(z - \mu_z)(z - \mu_z)'A'] \\ &= AE[(z - \mu_z)(z - \mu_z)']A' \\ &= AV_zA'. \end{aligned}$$

### B.2.1 Joint Distribution of Two Vectors

Now consider two random vectors, say  $\theta = (\theta_1, \dots, \theta_r)'$  and  $y = (y_1, \dots, y_n)'$  and the relationships between them. We will assume that the joint random vector  $(\theta', y')' = (\theta_1, \dots, \theta_r, y_1, \dots, y_n)'$  has a density function

$$p_{\theta,y}(u, v) \equiv p_{\theta,y}(u_1, \dots, u_r, v_1, \dots, v_n),$$

where  $(v_1, \dots, v_n)' \equiv v$  is a vector of numbers or placeholder variables used in association with  $y$ .

The distribution of one random vector, say  $\theta$ , ignoring the other vector,  $y$ , is called the *marginal distribution* of  $\theta$ . The marginal cdf of  $\theta$  can be obtained by putting  $+\infty$  into the joint cdf for all of the  $y$  variables:

$$F_\theta(u) = F_{\theta,y}(u_1, \dots, u_r, +\infty, \dots, +\infty).$$

The *marginal density* is obtained by either partial differentiation of  $F_\theta(u)$  or integrating the joint density over the  $y$  variables:

$$p_\theta(u) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p_{\theta,y}(u_1, \dots, u_r, v_1, \dots, v_n) dv_1 \cdots dv_n.$$

Denote the marginal density of  $y$ ,  $f_y(v)$ .

### B.2.2 Conditional Distributions

The conditional density of a vector, say  $\theta$ , given the value of the other vector, say  $y = v$ , is obtained by dividing the density of  $(\theta', y')'$  by the density of  $y$  evaluated at  $v$ , i.e.,

$$p_{\theta|y}(u|v) \equiv p_{\theta,y}(u, v) / f_y(v). \quad (1)$$

Similarly,

$$f_{y|\theta}(v|u) \equiv p_{\theta,y}(u, v) / p_\theta(u).$$

The conditional density (1) is a function of  $u$ , so the term  $f_y(v)$  in the denominator of the right-hand side is just a constant that makes the density integrate to 1. Anytime we have

$$p_{\theta|y}(u|v) \propto p_*(u)$$

for all  $u$ , where the constant of proportionality may depend on  $v$  but not  $u$ , we have

$$p_{\theta|y}(u|v) = \frac{p_*(u)}{\int p_*(u) du}$$

and we call  $p_*(u)$  a *kernel* of the conditional density.

Knowing  $f_{y|\theta}(v|u)$  and  $p_\theta(u)$  is, in principle, enough to let us find the conditional density  $p_{\theta|y}(u|v)$ . Unfortunately, it is often difficult to perform the calculus needed to find  $p_{\theta|y}(u|v)$ .

**Bayes' Theorem.** The density of  $\theta$  given  $y$  is

$$p_{\theta|y}(u|v) = \frac{f_{y|\theta}(v|u)p_\theta(u)}{\int f_{y|\theta}(v|u)p_\theta(u) du}$$

where the integral is over  $-\infty$  to  $\infty$  for every component of  $u$ .

**PROOF.** Observe that by the definition of a conditional density, the term in the numerator is

$$f_{y|\theta}(v|u)p_\theta(u) = \frac{p_{\theta,y}(u, v)}{p_\theta(u)} p_\theta(u) = p_{\theta,y}(u, v)$$

and the term in the denominator is just

$$\int f_{y|\theta}(v|u)p_\theta(u)du = f_y(v)$$

and so, the formula in Bayes' Theorem agrees with the definition (1). Essentially the same argument works if both distributions are either continuous or discrete. If one vector has a continuous distribution and the other is discrete, one needs to be a bit more careful with definitions but the result still holds.

Applications of Bayes' Theorem are given throughout the book but in particular see Sections 2.3 and 2.4. Note that  $f_{y|\theta}(v|u)p_\theta(u)$  is a kernel of the conditional density  $p_{\theta|y}(u|v)$ .

Using our convention of replacing placeholder variables with the symbols for the random variables, we typically write Bayes' Theorem as

$$p(\theta|y) = \frac{f(y|\theta)p(\theta)}{\int f(y|\theta)p(\theta)d\theta}.$$

Using the conditional density we can define conditional expectations. Let  $g$  be a function from  $\mathbf{R}^r$  into  $\mathbf{R}$ ,

$$E[g(\theta)|y=v] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(u)p_{\theta|y}(u|v)du,$$

where  $du \equiv du_1du_2 \cdots du_r$ . In particular, the conditional expectation of  $\theta_j$  given  $y$  is

$$E[\theta_j|y=v] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} u_j p_{\theta|y}(u|v)du$$

and, writing  $\hat{\theta}_j(v) \equiv E[\theta_j|y=v]$ , the conditional variance is

$$\begin{aligned} \text{Var}[\theta_j|y=v] &\equiv E\{[\theta_j - \hat{\theta}_j(v)]^2|y=v\} \\ &= \int \cdots \int [u_j - \hat{\theta}_j(v)]^2 p_{\theta|y}(u|v)du. \end{aligned}$$

Standard properties of expectations hold for conditional expectations. For example, with  $a$  and  $b$  real,

$$E[ag_1(\theta) + bg_2(\theta)|y=v] = aE[g_1(\theta)|y=v] + bE[g_2(\theta)|y=v].$$

The conditional expectation  $E[g(\theta)|y=v]$  is a function of the value  $v$ , but, since  $y$  is random, we can think of  $E[g(\theta)|y=v]$  as a random variable. In line with our policy of eliminating placeholder variables, write  $E[g(\theta)|y]$ . Similarly we write  $E[\theta_j|y]$  and  $\text{Var}[\theta_j|y]$ .

Thinking of conditional expectations as random variables, an important property is

### Proposition B.1

$$E[g(\theta)] = E\{E[g(\theta)|y]\}.$$

PROOF. To see this, note that  $p_{\theta|y}(u|v)f_y(v) = p_{\theta,y}(u,v)$  and

$$\begin{aligned} E\{E[g(\theta)|y]\} &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} E[g(\theta)|y=v] f_y(v) dv \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left[ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(u) p_{\theta|y}(u|v) du \right] f_y(v) dv \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(u) p_{\theta|y}(u|v) f_y(v) du dv \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(u) p_{\theta,y}(u,v) du dv \\ &= E[g(\theta)]. \end{aligned}$$

These results can be generalized to deal with functions  $g(\theta, y)$  from  $\mathbf{R}^{r+n}$  into  $\mathbf{R}$ . If  $y = v$ , it is obvious how to define  $E[g(\theta, y)|y = v]$ . If we consider  $y$  as random, write  $E[g(\theta, y)|y]$ . It is easily established that

$$E[g(\theta, y)] = E[E[g(\theta, y)|y]].$$

Note that a function of  $\theta$  or  $y$  alone can be considered as a function from  $\mathbf{R}^{r+n}$  into  $\mathbf{R}$ .

Somewhat more complex results hold for variances and covariances.

**Proposition B.2** Define a random variable  $\gamma \equiv g(\theta, y)$ , then

$$\text{Var}(\gamma) = \text{Var}[E(\gamma|y)] + E[\text{Var}(\gamma|y)].$$

Moreover, define  $\gamma_i \equiv g_i(\theta, y)$  for  $i = 1, 2$ , then

$$\text{Cov}(\gamma_1, \gamma_2) = \text{Cov}[E(\gamma_1|y), E(\gamma_2|y)] + E[\text{Cov}(\gamma_1, \gamma_2|y)].$$

A second important property of conditional expectations is

**Proposition B.3** If  $h(y)$  is a function from  $\mathbf{R}^n$  into  $\mathbf{R}$ , we have

$$E[h(y)g(\theta, y)|y] = h(y)E[g(\theta, y)|y].$$

PROOF. This follows because if  $y = v$ ,

$$\begin{aligned} E[h(y)g(\theta, y)|y = v] &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(v)g(u, v)p_{\theta|y}(u|v)du \\ &= h(v) \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(u, v)p_{\theta|y}(u|v)du \\ &= h(v)E[g(\theta, y)|y = v]. \end{aligned}$$

This is true for all  $v$ , so the proposition holds.

If  $g(\theta, y) \equiv 1$ , we get the special case

$$E[h(y)|y] = h(y).$$

Finally, we can extend the idea of conditional expectation to random vectors. Let  $g(\theta, y)$  be a function from  $\mathbf{R}^{r+n}$  into  $\mathbf{R}^s$ . Write  $g(\theta, y) = [g_1(\theta, y), \dots, g_s(\theta, y)]'$ . Then define

$$E[g(\theta, y)|y] = (E[g_1(\theta, y)|y], \dots, E[g_s(\theta, y)|y])'.$$

### B.2.3 Independence

If their densities exist, two random vectors are independent if and only if their joint density is equal to the product of their marginal densities, i.e.,  $\theta$  and  $y$  are independent if and only if

$$p_{\theta,y}(u, v) = p_{\theta}(u)f_y(v).$$

We use the notation

$$\theta \perp\!\!\!\perp y$$

to denote independence of  $\theta$  and  $y$ . If the random vectors are independent, then any vector valued functions of them, say  $g(\theta)$  and  $h(y)$ , are also independent. Note that if  $\theta$  and  $y$  are independent,  $p_{\theta|y}(u|v) = p_\theta(u)$ , or less formally,

$$p(\theta|y) = p(\theta),$$

i.e., the conditional density of  $\theta$  given  $y$  is the same as the marginal density of  $\theta$ . This is a more intuitive idea of independence. Intuitively, if  $\theta$  and  $y$  are independent,  $y$  should contain no information that would change our thinking about  $\theta$ , so the conditional distribution of  $\theta$  given  $y$  should be the same as the marginal distribution of  $\theta$ .

We can similarly define notions of conditional independence. Random vectors  $y_1$  and  $y_2$  are conditionally independent given  $\theta$  if

$$f(y_1, y_2 | \theta) = f_1(y_1 | \theta) f_2(y_2 | \theta).$$

We write this

$$y_1 \perp\!\!\!\perp y_2 | \theta.$$

#### B.2.4 Moment Generating Functions

For a random vector  $\theta = (\theta_1, \dots, \theta_r)'$ , the *moment generating function (MGF)* is a function from  $\mathbf{R}^r$  to  $\mathbf{R}$ . It is defined by

$$\varphi_\theta(t_1, \dots, t_r) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left[\sum_{j=1}^r t_j u_j\right] p_\theta(u_1, \dots, u_r) du_1 \cdots du_r.$$

We are interested in MGFs because if  $\theta = (\theta_1, \dots, \theta_r)'$  and  $y = (y_1, \dots, y_r)'$  are random vectors and if

$$\varphi_\theta(t_1, \dots, t_r) = \varphi_y(t_1, \dots, t_r)$$

for all  $(t_1, \dots, t_r)$ , then  $\theta$  and  $y$  have the same distribution.

#### B.2.5 Change of Variables

Consider transforming an  $n$  dimensional random vector  $z$  into a new  $n$  dimensional random vector  $y = G(z)$ . We need the transformation  $G$  to be invertible, so that  $G^{-1}(y) = z$ . If  $z$  has density  $f_z(v)$ , we want to find the density of  $y$ , say,  $q_y(u)$ . This involves transforming the density of  $z$  in an obvious way but it also involves multiplying by a term that accounts for the local change in coordinate system.

**Proposition B.4** The density of  $y = G(z)$  is

$$q_y(u) = f_z(G^{-1}(u)) |\det(dG^{-1}(u))|,$$

where  $dG^{-1}(u)$  is the derivative (matrix of partial derivatives) of  $G^{-1}$  evaluated at  $u$ .

The obvious way of transforming the density is through  $f_z(G^{-1}(u))$ . The term adjusting for local changes in the coordinate system is  $|\det(dG^{-1}(u))|$ .

Getting rid of placeholder variables, we will commonly write

$$q(y) = f(G^{-1}(y)) |\det(dG^{-1}(y))|.$$

**EXAMPLE B.1.** *Multivariate normal.* Let  $z_1, \dots, z_n$  be iid  $N(0, 1)$ . Write the random vector  $z = (z_1, \dots, z_n)'$ . Because the  $z_i$ s are  $N(0, 1)$ , their (marginal) densities are  $f_*(z_i) = (1/\sqrt{2\pi}) \exp[-z_i^2/2]$ .

By independence, their joint density is

$$\begin{aligned} f(z) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-z_i^2/2} \\ &= \left( \frac{1}{\sqrt{2\pi}} \right)^n e^{-\sum_{i=1}^n z_i^2/2} \\ &= \left( \frac{1}{\sqrt{2\pi}} \right)^n \exp \left[ \frac{-1}{2} z' z \right]. \end{aligned}$$

Now we consider a transformation  $y = G(z) \equiv Az + \mu$  where  $A$  is a fixed  $n \times n$  nonsingular matrix and  $\mu$  is a fixed  $n$  vector. Before proceeding with the transformation, note that  $E(z)$  is a vector of 0s and  $\text{Cov}(z) = I$ . Also, just from using linearity properties of expectations,

$$E(y) = E(Az + \mu) = AE(z) + \mu = A0 + \mu = \mu$$

and

$$\text{Cov}(y) = A\text{Cov}(z)A' = AA'.$$

With  $y = Az + \mu$ , the inverse transformation is  $z = G^{-1}(y) = A^{-1}(y - \mu)$ . It is not difficult to see that the derivative of the inverse transformation  $G^{-1}(y)$  is  $dG^{-1}(y) = A^{-1}$ . Thus, the density of  $y$  is

$$\begin{aligned} q(y) &= f(A^{-1}(y - \mu)) |\det(A^{-1})| \\ &= \left( \frac{1}{\sqrt{2\pi}} \right)^n \exp \left\{ \frac{-1}{2} [A^{-1}(y - \mu)]' [A^{-1}(y - \mu)] \right\} |\det(A^{-1})| \\ &= \left( \frac{1}{\sqrt{2\pi}} \right)^n \exp \left\{ \frac{-1}{2} (y - \mu)' A^{-1}' A^{-1} (y - \mu) \right\} |\det(A^{-1})|. \end{aligned}$$

We can put the density in a nicer form by denoting the covariance matrix  $V = AA'$  so that  $V^{-1} = A^{-1}' A^{-1}$ . Using properties of determinants

$$\det(A^{-1}) = \sqrt{\det(A^{-1}' A^{-1})} = \sqrt{\det(V^{-1})} = 1/\sqrt{\det(V)}.$$

Incorporating these facts, the density can be rewritten as

$$q(y) = \left( \frac{1}{\sqrt{2\pi}} \right)^n \left( \frac{1}{\sqrt{\det(V)}} \right) \exp \left\{ \frac{-1}{2} (y - \mu)' V^{-1} (y - \mu) \right\}.$$

Note that the density, and thus the distribution, depends on  $A$  only through the covariance matrix  $V = AA'$ . We write this distribution as

$$y \sim N_n(\mu, V).$$

It is a simple matter to show that linear transformations of multivariate normal distributions are again multivariate normal. In particular, if  $y \sim N_n(\mu, V)$ , then for fixed matrix  $T$  and vector  $u$ ,

$$Ty + u \sim N(T\mu + u, TVT').$$

### B.3 Models and Conditional Independence

Throughout this book, we most often specify joint probability models for  $y$  and  $\theta$  by specifying the conditional distribution for  $y$  given  $\theta$ , perhaps through a density  $f(y|\theta)$ , and then the marginal

distribution of  $\theta$ , with density  $p(\theta)$ . The density notation is convenient for general discussion but inconvenient for specific examples. In particular, we might specify

$$y|\theta \sim \text{Bin}(n, \theta)$$

and

$$\theta \sim \text{Beta}(\alpha_1, \alpha_2).$$

Less formally, we might simply specify

$$y \sim \text{Bin}(n, \theta)$$

where it is understood that the distribution of  $y$  is conditional on  $\theta$  because  $\theta$  appears as a parameter in the binomial distribution. The distribution is also conditional on  $n$  but we typically treat  $n$  as known in binomial sampling. Similarly, we might specify the model

$$y|\theta \sim N(\theta, 1)$$

and

$$\theta \sim N(\alpha_1, \alpha_2).$$

Again, if we specify the model

$$y \sim N(\theta, 1),$$

it is understood that the distribution of  $y$  is conditional on  $\theta$  because  $\theta$  appears as a parameter in the normal distribution. In general, we specify

$$y|\theta \sim f(y|\theta)$$

and

$$\theta \sim p(\theta)$$

with an informal specification of the conditional distribution as

$$y \sim f(y|\theta).$$

Sometimes things get a bit more complicated. Suppose we have  $n$  pairs of identical twins and one twin is randomly selected to get a drug  $D_1$  and the other gets  $D_2$ . Some original response is measured on each twin, but the only data remaining are the differences in the twins' responses, say,

$$y_1, \dots, y_n \text{ iid } N(\gamma, 1)$$

with density  $f(y|\gamma)$ . It is quite plausible that there would be no direct information available to determine a (prior) distribution for  $\gamma$ ; however, there should be information available on the drugs individually, say on  $\theta_1$ , the mean response for people on  $D_1$  and also for  $\theta_2$ , the mean response for  $D_2$ . This would give us a joint distribution on  $\theta_1$  and  $\theta_2$  with, say, density  $p(\theta_1, \theta_2)$ . From this joint density, we could derive a density for  $\gamma \equiv \theta_1 - \theta_2$ , say,  $q(\gamma)$ , which in turn determines the joint density  $p(y, \gamma) = f(y|\gamma)q(\gamma)$ . (In this scenario with identical twins the variability in  $q(\gamma)$  would probably be inappropriately high.)

For reasons that will only become clear later, we sometimes need to discuss the joint density  $p(y, \theta_1, \theta_2)$ . In particular, we want to be able to conclude that  $p(y, \theta_1, \theta_2) = f(y|\theta_1 - \theta_2)p(\theta_1, \theta_2)$ . In this case, it seems clear that aspects of the joint distribution of  $\theta_1$  and  $\theta_2$  other than the marginal distribution of  $\theta_1 - \theta_2$  are irrelevant to the distribution of  $y$ .

We now expand on this example to present a general discussion of model specification and conditional independence that is designed as a theoretical background for the discussion of hierarchical models in Section 4.12. It might be better to read all of Section 4.12 before wading into this morass.

It would certainly be good to read the four examples in Section 4.12 to get some idea of specific problems that need to be addressed.

In general, let  $y$  be an  $n$  vector,  $\theta$  an  $r$  vector, and  $\gamma(\theta)$  a vector valued function. A key distinction is that  $y$  is observable and that  $\theta$  is not observable. Ultimately, we will want the conditional distribution of the unobservable  $\theta$  (or some function of it) given the observable  $y$ . Let the density on  $\theta$  be  $p(\theta)$ . As with the identical twins, it is sometimes convenient to specify the distribution of the observables as

$$y|\gamma(\theta) \sim f(y|\gamma(\theta)).$$

In such cases we will implicitly assume that

$$y \perp\!\!\!\perp \theta | \gamma(\theta).$$

In words, we are assuming that the observables are independent of the unobservables given the parameter used to specify the conditional distribution of the observables. With this assumption, first by conditional independence and then by the fact that knowing  $\theta$  implies knowledge of  $\gamma(\theta)$ ,

$$y|\gamma(\theta) \sim y|\gamma(\theta), \theta \sim y|\theta.$$

It follows that

$$f(y|\gamma(\theta)) = f(y|\theta)$$

and the joint density is

$$p(y, \theta) = f(y|\theta)p(\theta) = f(y|\gamma(\theta))p(\theta).$$

Less formally, we might specify the conditional distribution of  $y$  as

$$y \sim f(y|\gamma(\theta)).$$

One special case of particular importance is the prediction of future observables  $\tilde{y}$ . They are currently unobservable, so if we have a sampling distribution  $y \sim f(y|\theta)$ , unless otherwise stated, we implicitly assume that  $y \perp\!\!\!\perp \tilde{y} | \theta$ .

It is also sometimes convenient to specify the distribution on  $\theta$  via conditional distributions. Suppose we decompose the vector  $\theta$  into three subvectors,  $\theta = (\theta_0', \theta_1', \theta_2')'$ . We might specify the density on  $\theta$  as

$$p(\theta_0, \theta_1, \theta_2) = p_0(\theta_0|\theta_1)p_1(\theta_1, \theta_2).$$

It is not clear that this defines a valid joint density. A valid joint density is

$$p(\theta_0, \theta_1, \theta_2) = p_0(\theta_0|\theta_1, \theta_2)p_1(\theta_1, \theta_2).$$

However, if we assume that  $\theta_0 \perp\!\!\!\perp \theta_2 | \theta_1$ , we get

$$p_0(\theta_0|\theta_1) = p_0(\theta_0|\theta_1, \theta_2).$$

Whenever we specify a conditional distribution for one set of unobservables ( $\theta_0$ ) given another set of unobservables ( $\theta_1$ ), we will implicitly assume that the first set is independent of the remaining unobservables ( $\theta_2$ ) given the conditioning unobservables. Not infrequently, these results are applied after a reparameterization of  $\theta$  into, say,  $\tilde{\gamma} = [\gamma(\theta)', \gamma_1(\theta)', \gamma_2(\theta)']'$ .

Finally, decompose  $\theta$  into two parts  $\theta_0$  and  $\theta_1$ . If we specify a sampling distribution

$$y|\theta_0 \sim f(y|\theta_0),$$

clearly the data only give us information about  $\theta_0$ . That is reflected in the fact that the posterior distribution of  $\theta$  given  $y$  can be decomposed into

$$p(\theta|y) = p_0(\theta_0|y)p_1(\theta_1|\theta_0)$$

where  $p_0(\theta_0|y)$  is the standard posterior based on the marginal prior for  $\theta_0$  (ignoring  $\theta_1$ ) and the conditional density  $p_1(\theta_1|\theta_0)$  is determined entirely by the prior  $p(\theta_0, \theta_1)$ . We now demonstrate this fact.

With our standard assumptions for conditional specification of distributions for observables,  $y \perp\!\!\!\perp \theta_1 | \theta_0$ , so

$$f(y|\theta) = f(y|\theta_0).$$

The prior  $p(\theta)$  can be rewritten as  $p_1(\theta_1|\theta_0)p_0(\theta_0)$ . Applying Bayes' Theorem

$$\begin{aligned} p(\theta|y) &= \frac{f(y|\theta)p(\theta)}{f(y)} \\ &= \frac{f(y|\theta_0)p(\theta_1|\theta_0)p_0(\theta_0)}{f(y)} \\ &= p(\theta_1|\theta_0) \frac{f(y|\theta_0)p_0(\theta_0)}{f(y)} \\ &= p(\theta_1|\theta_0)p_0(\theta_0|y). \end{aligned}$$

This result assumes major importance in the discussion of identifiability in Section 4.14.

**EXERCISE B.1.** Let  $\theta$  and  $y$  be independent. Show that

- (a)  $E[g(\theta)|y] = E[g(\theta)]$ .
- (b)  $E[g(\theta)h(y)] = E[g(\theta)]E[h(y)]$ .

**EXERCISE B.2.** Let  $y$  have pdf  $f(u) = 3u^2I_{(0,1)}$ . Define  $x = \sqrt{y}$ . Find the pdf of  $x$ ,  $E(x)$ , and  $\text{Var}(x)$ .

**EXERCISE B.3.** Relative to Proposition B.4, show that

$$|\det(dG^{-1}(u))| = 1/|\det\{dG[G^{-1}(u)]\}|.$$

Hint: This follows from three facts: (a) for two nonsingular matrices  $\det(AB) = \det(A)\det(B)$ , (b)  $G(G^{-1}(u)) = u$ , and (c) the chain rule.



---

# Appendix C: Getting Started in R

---

This appendix gives a brief overview of the R language for statistical computing and data presentation, cf. Ihaka and Gentleman (1996) and R Development Core Team (2008). The main benefits of R include the price (it's free!), its graphical capabilities, its programmability (users can program their own functions), and its open-source environment. Regarding the latter, R users can package together functions and provide them to the entire R community through the Comprehensive R Archive Network (CRAN; <http://cran.r-project.org/>). These packages, called R libraries, provide, in a timely manner, routines for implementing state-of-the-art statistical methods. However, a trade-off for quick dissemination is that the code is probably less thoroughly examined than commercial software.

The CRAN website contains many valuable resources (including an extensive users manual) that provide details beyond those presented here. Various copies ("mirrors") of the CRAN website exist around the world. Books devoted to statistical data analysis using R (or its commercially available progenitor S-PLUS) include Venables and Ripley (2002), Krause and Olson (2005), Everitt and Hothorn (2006), Crawley (2007), and Dalgaard (2008). The text by Albert (2007) is devoted entirely to Bayesian computing with R. Moreover, there are quite a number of R packages devoted to Bayesian analysis, cf.,

<http://cran.r-project.org/web/views/Bayesian.html>.

The FEV data discussed in Chapters 7 and 9 and the leukemia survival data from Chapter 12 are used as illustrations here. The R code and data used in this book are available at our website, <http://www.stat.unm.edu/~fletcher/>.

## C.1 Getting R

Version 2.7.1 of R for Windows was used in this book. Shortly before the book went to press, version 2.10.1 was released. R is available at no cost from either the CRAN website or from <http://www.r-project.org/>. At the CRAN website, first select an operating system (e.g., Macintosh or Windows). Windows users will select base, then select Download R 2.10.1 for Windows, and finally Run the executable file to install R. Updated versions of R appear regularly. Check for them at the websites.

## C.2 Some R Basics

- R commands can be entered at the prompt (>), but we prefer to type commands into a text file (for instance, a Notepad file) and copy them into R.
- R is case sensitive.
- The comment symbol is #. Any text on a line after this symbol is ignored by R.
- Assignment operators include a (three keystroke) backarrow and the equals sign, i.e., <- and =. The concatenate command (c()) creates vectors using these operators.

```
y <- c(44,15,87,-4)
x = c(pi,exp(1),sqrt(3),10^2)
z <- c(rep(4,5),16,NA)
```

In `x` we are using the irrational number  $\pi$  and the function values  $\exp(1) \equiv e^1$ ,  $\sqrt{3}$ , and  $10^2$ . Syntactical features in `z` will be explained soon.

- Entering the name of an object at the R prompt prints out its contents:

```
> x
[1] 3.141593 2.718282 1.732051 100.000000
```

To access the second element of `x`, type

```
> x[2]
[1] 2.718282
```

The vector `z` includes a missing data point `NA` and the function `rep(4, 5)`, where `rep` is short for repeat. Here `rep(4, 5)` causes the number 4 to occur five times. In general, `rep(a, b)` creates a vector containing the value `a` repeated `b` times.

```
> z
[1] 4 4 4 4 4 16 NA
```

Other tricks useful in defining vectors are discussed in Section C.5.

- R includes many built-in scalar functions such as the square root (`sqrt`) and exponential (`exp`) functions used earlier. Scalar functions are applied to vectors elementwise, e.g.,

```
> sqrt(z)
[1] 2.0000 2.0000 2.0000 2.0000 2.0000 4.0000 NA
```

- Standard statistics are obtained by applying built-in functions to a vector of data values:

```
mean(x)
median(x)
var(x)
sd(x)
range(x)
min(x)
max(x)
length(x)
sum(x)
sort(x)
quantile(x, c(0.25, 0.75))
#rm(x)
```

Most of these are self-explanatory. The penultimate command finds first and third quartiles, i.e., the 25th and 75th percentiles of the empirical distribution function (edf) based on the data in `x`. In general, `quantile(x, y)` finds the percentiles listed in `y` for the data (edf) in `x`. The last command in this list would remove `x` from the session, but would not be executed because it is commented out.

- The default setting in many R functions is to return an error or `NA` when applied to a vector that contains missing values. For instance,

```
> mean(z)
[1] NA
```

To compute the mean of the observed values in `z` by removing the `NA` values, use

```
> mean(z, na.rm=TRUE)
[1] 6.000000
```

Options on how to handle missing data are commonly provided by the function arguments `na.rm` or `na.action`. Details are provided in the help file for any particular R function.

- To access a help file for a known function, say `mean`, enter `?mean` or `help(mean)` at the R prompt. This particular help file shows that the default is `na.rm=FALSE` and that trimmed means can be calculated. If you don't know the name of a function, use `help.search`, e.g.,

`help.search("regression")`, or simply do an online search for help. R users can subscribe online to the R mailing list or search through an archive of previous questions and answers submitted by others (<http://tolstoy.newcastle.edu.au/R/>).

- Matrix operations are transpose (`t()`), inverse (`solve()`), and matrix multiplication (`%*%`). Scalar multiplication uses the symbol (`*`).

First define a matrix `x1` from the vector `x`.

```
> x1 <- matrix(x,nrow=2,ncol=2)
> x1
     [,1]      [,2]
[1,] 3.141593  1.732051
[2,] 2.718282 100.000000
```

By default R fills a matrix columnwise.

```
> y1 <- matrix(y,2,2)
> y1
     [,1] [,2]
[1,]   44   87
[2,]   15   -4
```

The well-known `Vec` operator would turn the matrices `x1` and `y1` back into the original vectors `x` and `y`.

Applying the transpose function gives

```
> t(x1)
     [,1]      [,2]
[1,] 3.141593  2.718282
[2,] 1.732051 100.000000
```

To find the inverse of `x1`, apply

```
> solve(x1)
     [,1]      [,2]
[1,] 0.323152873 -0.005597172
[2,] -0.008784206  0.010152147
```

Finally, to multiply `x1` and `y1`,

```
> x1%*%y1
     [,1]      [,2]
[1,] 164.2108 266.3904
[2,] 1619.6044 -163.5095
```

- The command `x1[2, 1]` returns 2.718282, the element in row 2, column 1 of `x1`. The command `x1[2, ]` returns the vector containing the second row of `x1`: 2.718282 and 100.000000. Similarly, `x1[, 1]` gives the first column of `x1`. We can apply functions to these vectors just as before, e.g., `mean(x1[2, ])` returns 51.35914.
- We can also apply a function to each row or each column of a matrix. Consider the task of computing the mean value for each column in a matrix. This is useful when standardizing regression predictor variables or when Gibbs sampler iterates for different parameters are stored in columns of a matrix. In the latter case, the mean of the columns gives an approximation to the posterior mean vector. The mean of each column of `x1` and the minimum value in each row of `x1` are obtained below. The second argument of the `apply` function indicates whether the operation is to be applied to rows (1) or to columns (2).

```
> apply(x1,2,mean)
[1] 2.929937 50.866025
```

```
> apply(x1,1,min)
[1] 1.732051 2.718282
```

- Density, cumulative distribution, and quantile functions are available for many standard distributions. For normal distributions, all three functions have a first argument that is a scalar or vector where the functions are evaluated with second and third arguments that are the mean and standard deviation of the normal. For example, the code below evaluates the  $N(0, 1)$  density at 0 and the cdf at 1.645. It also returns the 2.5th, 50th, and 97.5th percentiles of the distribution.

```
> dnorm(0,mean=0,sd=1)
[1] 0.3989423
> pnorm(1.645,0,1)
[1] 0.950015
> qnorm(c(0.025,0.50,0.975),0,1)
[1] -1.959964  0.000000  1.959964
```

R parameterizes the normal using the standard deviation instead of the variance or precision. Similar functions are available for other distributions, including the Poisson, uniform, binomial,  $t$ , and gamma (e.g., `dpois(x,lambda)`, `dunif(x,min,max)`, `dbinom(x,size,prob)`, `dt(x,df)`, and `dgamma(x,shape,rate)`). Like WinBUGS, R uses the parametrization for  $\text{Gamma}(a,b)$  with mean  $a/b$  and variance  $a/b^2$ .

Random samples from any of these distributions are simulated by using “r” as the first letter, for example, `rnorm(1000,4,5)` generates a random sample of 1000 values from the normal distribution with mean 4 and variance 25.

### C.3 User-Contributed Packages

In addition to built-in functions provided in its base package, R enables users to write their own functions and store them in libraries that can be downloaded through CRAN. For instance, functions for MCMC convergence diagnostics are available in the `coda` (convergence diagnosis and output analysis) and `boa` (Bayesian output analysis) libraries. The library `MCMCpack` contains functions for fitting certain models (e.g., multinomial logistic regression) via MCMC sampling and it also provides useful functions (e.g., for simulating from Dirichlet and Wishart distributions) for specialized programming. The library `DPPackage` contains functions for fitting some of the Bayesian nonparametric models discussed in Chapter 15. To see a complete list and description of the functions contained in a particular library, type `library(help=name)` where `name` is the library name (e.g., `library(help=DPPackage)`).

As an example, consider the `Hmisc` library, which contains several functions for exploratory data analysis. Its `summarize` function provides descriptive statistics stratified by groups. First we need to download and install this package, which is accomplished using `install.packages("Hmisc")` or through the toolbar in R (go to Packages, select Install Package(s) . . . , select a CRAN mirror near you to see a complete list of available libraries, and scroll down the alphabetical list until `Hmisc` is found). After importing the FEV data into R (details in the next section), the following code calculates the mean and standard deviation of FEV separately for smokers and nonsmokers.

```
> library(Hmisc)
> summarize(FEV,by=Smoke,FUN=mean)
  Smoke      FEV
1     0 3.150420
2     1 3.297547
> summarize(FEV,by=Smoke,FUN=sd)
  Smoke      FEV
1     0 0.7590191
2     1 0.7369897
```

#### C.4 Reading Data

Data are commonly stored in text files, Excel files, or at web sites. We illustrate methods for working with all three types using the FEV data. Of course, only one type would generally be used for a data set.

```
FEVdata1 <- read.table("FEVdata.txt", header=T, sep="\t")
attach(FEVdata1)
FEVdata1
summary(FEVdata1)
```

- Including `header=T` in `read.table` tells R that the first row of the text file contains variable names. If names are not included in row 1, we must define them by adding `, col.names=c("Age", "FEV", "Smoke")` to the `read.table` function.
- In `read.table`, the (default) command `sep=" "` indicates that a blank space is used to separate data values. Similarly, `sep="\t"` indicates that data values in each row are separated by tabs. For comma separated values, use `sep=","`. The command `sep=""` seems to read correctly most data files that use spaces as separators.
- The required `attach` command allows operations to be performed on the individual variables (in this case, Age, FEV, and Smoke). The result of `read.table` and `attach` is a matrix `FEVdata1` with labeled columns.
- We recommend printing the data, or a subset of them for very large data sets, and scanning through them visually to ensure that they were read correctly. Do this by typing the name given to the data, e.g., `FEVdata1` or `print(FEVdata1)`. To print just the first 10 rows enter `FEVdata1[1:10, ]`.
- The `summary` command gives some descriptive statistics for each variable in the data.
- The `read.table` command as presented above assumes that the data file is in the same directory (folder) in which R is being run. More generally, the first argument would be `"c:\\\\data\\\\FEVdata.txt"` if the data file is in the folder data on drive c. If the file is on the internet, R has the capability of reading it directly, e.g., `read.table("http://anson.ucdavis.edu/~johnson/FEVdata.txt", col.names=c("Age", "FEV", "Smoke"))`.

To import Excel 97 or Excel 2003 files, use the `read.xls` function contained in the `xlsReadWrite` library. See the function `odbcConnect` in the `RODBC` library for details on handling other file types, including data stored in newer versions of Excel. Below are commands for reading Excel files.

```
library(xlsReadWrite)
FEVdata2 <- read.xls("FEVdata.xls", colNames=TRUE)
attach(FEVdata2)
FEVdata2
summary(FEVdata2)
```

Including `colNames=TRUE` in `read.xls` tells R that the first row of the Excel file contains variable names. If names are not included in row 1, then we define them using `colNames=c("Age", "FEV", "Smoke")`.

#### C.5 Graphing

A primary advantage of R is the ease with which top-notch graphics can be produced. Here we illustrate scatterplots, boxplots, and histograms. We show how to put multiple objects in a single plot and how to plot multiple graphs on one page. The examples build on one another, introducing and amplifying common elements of the graphical procedures.

To produce plots suitable for inclusion in LATEX documents, first run code to produce a graph. From the toolbar go to the **File** menu and select **Save as**. There are several options, postscript is one (pdf is another). If postscript is selected, then one can further select between .ps and .eps files.

- Figure C.1 presents a scatterplot of FEV and Age with two LOWESS nonparametric regression curves, one for smokers and one for nonsmokers. It was produced using the following code.

```
plot(Age,FEV)
lines(lowess(Age[Smoke==1],FEV[Smoke==1]),lwd=2)
lines(lowess(Age[Smoke==0],FEV[Smoke==0]),lty=2,lwd=2)
legend("bottomright",c("S","NS"),lty=c(1,2))
```

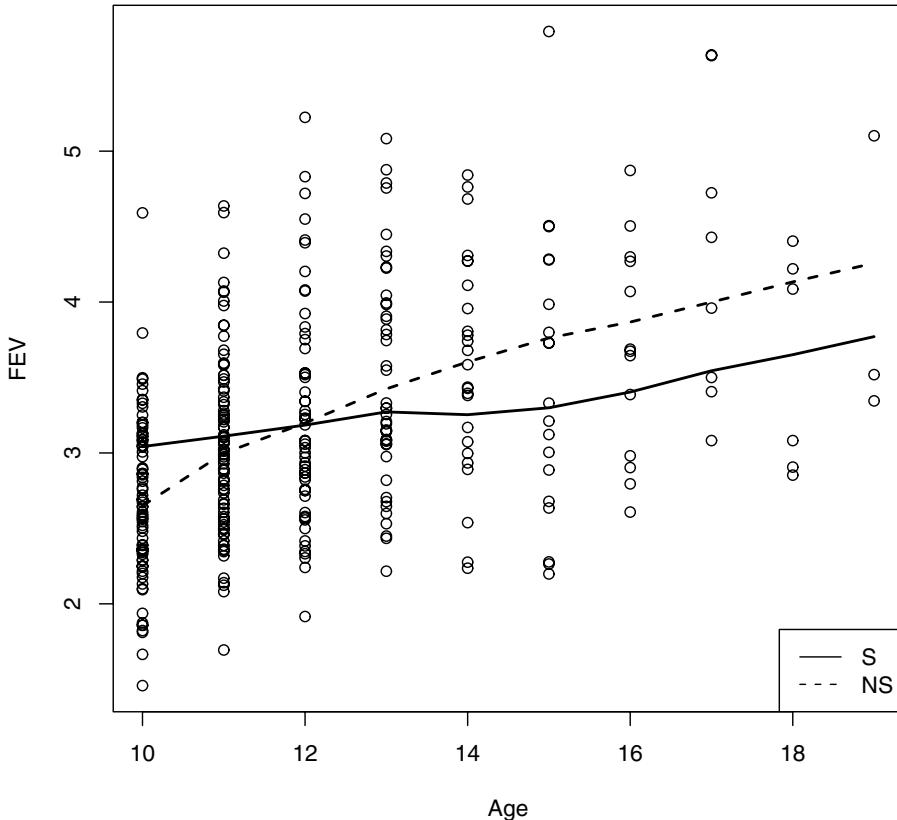


Figure C.1: Scatterplot of FEV with LOWESS curves.

The `plot` command is self-explanatory. The `lines` command is used to add additional information to a basic command like `plot` or `boxplot`. The `lowess` command takes two input vectors, `Age` and `FEV`, and produces two vectors: one a grid of Ages and the other the corresponding fitted `FEV` values. As applied here, `Age[Smoke==1]` specifies a vector of Ages that corresponds only to smokers, i.e., `Smoke = 1`. The `lines` command adds a graph of the LOWESS output vectors, which by default connects the dots with a solid line rather than plotting individual points. The argument `lwd` (line width) controls the thickness of the line. The argument `lty=2` changes the line type from the default solid line (`lty=1`) to a dashed line. The `legend` command is used to distinguish the line for smokers (S) from nonsmokers (NS).

- Our next example presents code for plotting the  $\text{Gamma}(5, 2)$  density, cf. Figure C.2. The function `seq(a, b, c)` produces a vector ranging from `a` to `b` in steps of length `c`.

```
x <- seq(0.01, 10, 0.01)
plot(x, dgamma(x, 5, 2), type="l", ylab="f(x)", main="Gamma(5, 2)")
```

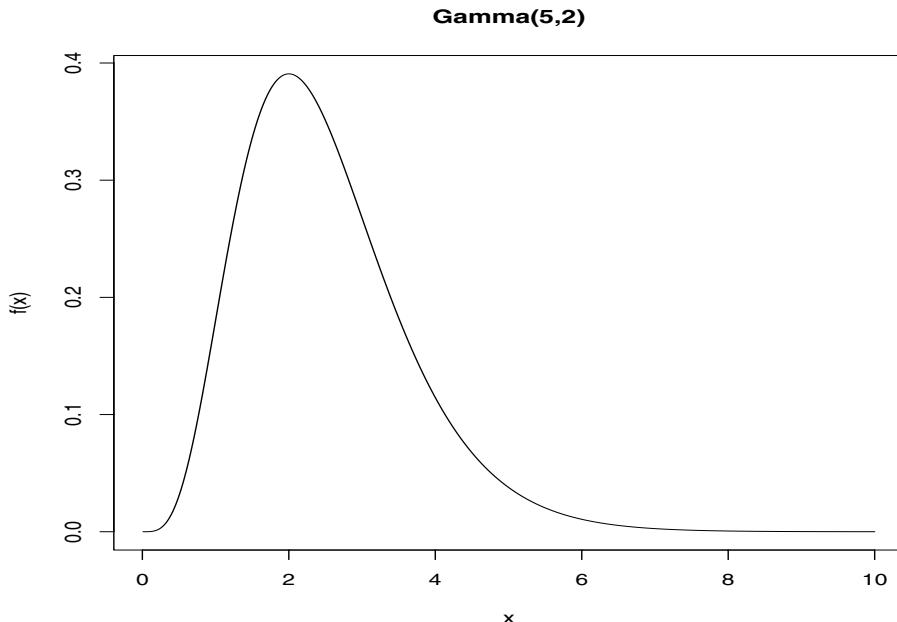


Figure C.2: *Gamma(5, 2) density.*

By default, `plot` produces a graph with circles for each data point. The option `type="l"` instructs `plot` to connect the dots to produce a curve. In `type="l"`, the argument is an  $\ell$  not the number 1. The commands `ylab` and `main` label the vertical axis and the overall graph, respectively. The horizontal axis label `x`, is the default value, the first argument of `plot`. This is changed using `xlab=" "` and inserting the desired label between the quotes.

- Figure C.3 gives side-by-side boxplots with accompanying line plots of the FEV data for smokers and nonsmokers. It was created using the following R code:

```
g=factor(Smoke,c(0,1),c("Nonsmoker","Smoker"))
boxplot(FEV~g,range=0,boxwex=0.25,ylab="FEV",xlab="")
lines(rep(1.2,length(FEV[Smoke==0])),FEV[Smoke==0],type="p")
lines(rep(2.2,length(FEV[Smoke==1])),FEV[Smoke==1],type="p")
```

The first line of code creates a group variable `g` where groups are distinguished by the 0/1 levels of `Smoke` and descriptive labels are provided. In general, `factor` is used to create categorical variables in R. As a result, arithmetic operations cannot be applied to them (e.g., try to run `mean(g)` and see what you get). The main components of the general syntax `factor(x, levels, labels)` include a data vector `x` that usually contains only a few distinct elements, a `levels` vector that specifies the symbols used for those distinct elements, and a character vector that labels the levels. Any value in `x` that does not correspond to a value in `levels` appears as a missing value in the output vector. By default `levels` identifies all the unique values in `x`, i.e., `levels=sort(unique.default(x))`. To invoke the default, use `g=factor(Smoke, labels=c("Nonsmoker", "Smoker"))`.

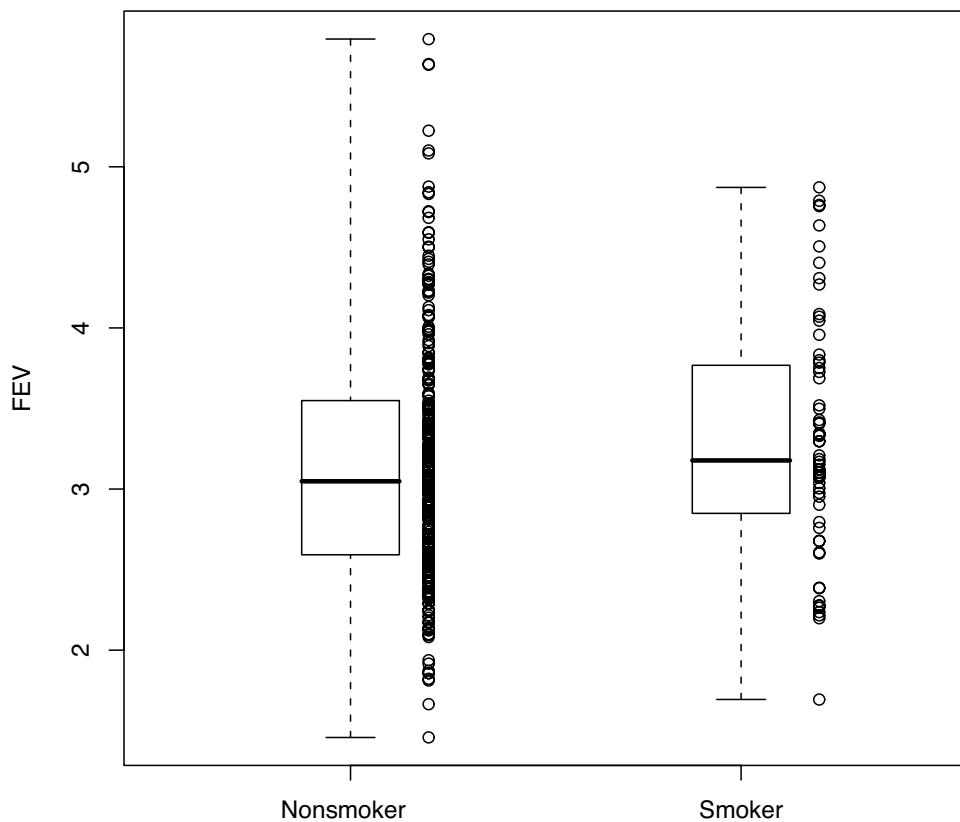


Figure C.3: Boxplots of the FEV data.

In the `boxplot` command, the first argument determines that a separate FEV boxplot is constructed for each level of `g`. The option `range=0` results in the whiskers being extended to the minimum and maximum FEV values, the option `boxwex=0.25` defines the width of the box, and `xlab` and `ylab` enable the user to specify labels for the x-axis and y-axis, respectively. Here the x-axis label is left blank.

The two `lines` commands produced the vertical line plots of the data. These are actually scatterplots superimposed on the boxplot graph. The basic command is `lines(x, y)`, a function of two equal length vectors. In our first application, the pairs of points are `(1.2, FEV)` where the FEV values are for nonsmokers only. The command `FEV[Smoke==0]` defines a vector of FEV values for nonsmokers (`Smoke = 0`) and `rep(1.2, length(FEV[Smoke==0]))` defines a vector of the same size, i.e., the number of nonsmokers, all entries having the value 1.2. The option `type="p"` instructs `lines` to plot individual data points rather than the default of plotting a line (`type="l"`).

- Histograms and kernel density estimates are obtained using `hist` and `density`. The following code produces the four plots in Figure C.4. To place them on a single graphic we use `par(mfrow=c(2,2))`, which is read “partition multiple figures by row into a  $2 \times 2$  page layout.”

```
par(mfrow=c(2,2))

hist(FEV,main="",ylim=c(0,100))    # Plot 1

hist(FEV,main="",prob=T,ylim=c(0,0.6))
```

```

lines(density(FEV))      # Plot 2

plot(density(FEV[Smoke==0]),type="l",xlab="FEV",
     ylab="Density", main="All Ages", xlim=c(1,7),ylim=c(0,0.7))
lines(density(FEV[Smoke==1]),lty=2)
legend("topright",c("NS","S"),lty=c(1,2))   # Plot 3

plot(density(FEV[Smoke==0 & Age>=14]),type="l",xlab="FEV",
     ylab="Density", main="Age 14 and Older",
     xlim=c(1,7),ylim=c(0,0.7))
lines(density(FEV[Smoke==1 & Age>=14]),lty=2)
legend("topright",c("NS","S"),lty=c(1,2))   # Plot 4

```

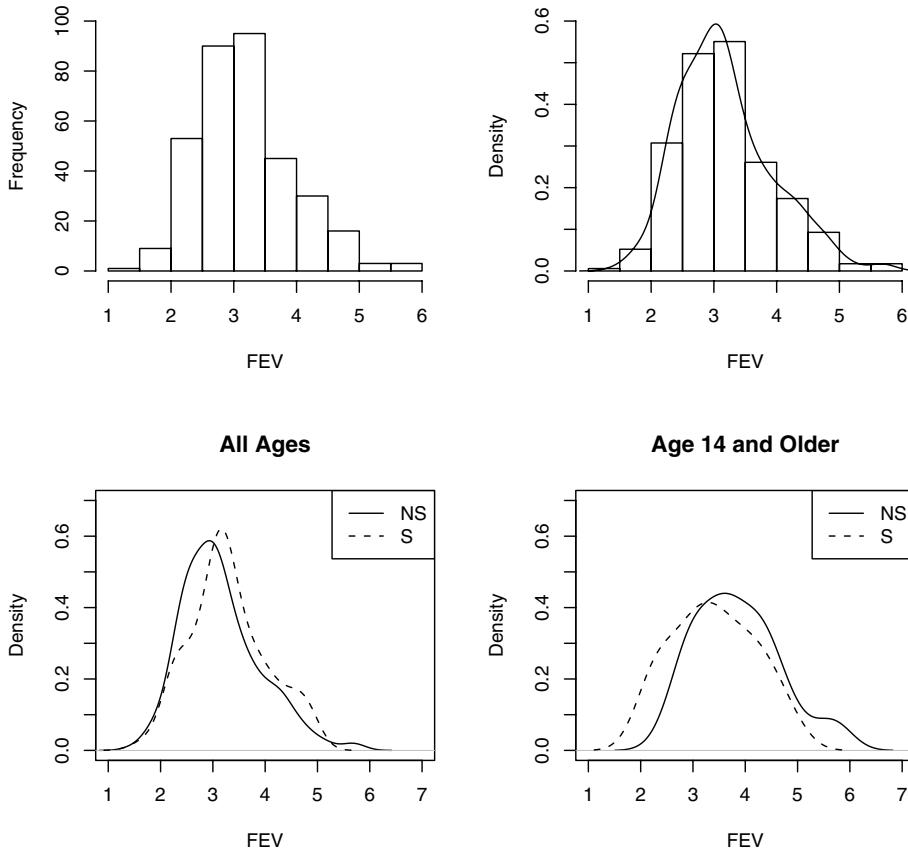


Figure C.4: Density estimates of FEV. S = smoker and NS = nonsmoker.

The first plot is a frequency histogram of the FEV data. By using `main=""`, we suppress a title from being displayed above the histogram and `ylim=c(0,100)` forces the y axis to range from 0 to 100.

In plot 2 (top right panel of Figure C.4), a probability histogram (`prob=T`) is presented together with a kernel density estimate of the FEV distribution. The probability histogram is rescaled so

that the area of the histogram is 1. The command `density` is similar to `lowess`; it takes the input `FEV` and produces both a grid of possible `FEV` values and corresponding density estimates.

The last two plots contain kernel density estimates of the distributions of `FEV` for smokers and nonsmokers using the complete data (plot 3; bottom left panel) and using the subset of data from  $\text{kids} \geq 14$  years of age (plot 4; bottom right panel). Curves rather than data points are plotted because `type="l"`. Different line types are produced by changing `lty`. Labels and legends as well as main titles are incorporated. Two logical restrictions are placed on `FEV` values in plot 4; `FEV` values with `Smoke = 0` and  $\text{Age} \geq 14$  are specified by `FEV[Smoke==0 & Age>=14]`.

When placing multiple plots into a graphic, one often wishes to use either the same vertical or horizontal axes. The commands `ylim` and `xlim` enable this. Typically, we examine the figures without these commands and then incorporate them as needed. Sometimes R will not show the entire graph without their use. Try running plot 3 without `ylim=c(0,0.7)`.

Note that R commands can span multiple lines.

Our final illustrations of graphics come from the analysis of the Feigl and Zelen (1965) data of Chapter 12 and presume familiarity with that chapter and WinBUGS. The data came from a comparative study of the survival time in weeks of two groups of patients (AG positive and AG negative) who died of leukemia. As in Chapter 12, we consider an  $\text{Exp}(\theta_1)$  sampling model for data  $y = (y_1, \dots, y_{n_1})$  from AG positive patients (Group 1) and an  $\text{Exp}(\theta_2)$  model for data  $x = (x_1, \dots, x_{n_2})$  from AG negative patients (Group 2). Our ultimate goal is to obtain Bayesian estimates of the survival functions  $S_1(t) = e^{-\theta_1 t}$  and  $S_2(t) = e^{-\theta_2 t}$  for all  $t > 0$ . In general, for a time to event variable  $T$  that varies according to a distribution with cdf  $F$ , the survival function is  $S(t) = P(T > t) = 1 - F(t)$ .

There are two ways to use WinBUGS to carry out the computations. A brute force method uses the “coda” option located in the WinBUGS Sample Monitor Tool to output all the simulated values, say  $\theta_1^1, \dots, \theta_1^{50000}$ , which can be saved like any data file and read into R for manipulation. A better option is to drive WinBUGS directly from within R as illustrated in the next section. In either case, we need a WinBUGS model:

```
model
{
  for(i in 1:n1){
    y[i] ~ dexp(theta1)
  }
  for(i in 1:n2){
    x[i] ~ dexp(theta2)
  }
  theta1 ~ dgamma(0.01,0.01)
  theta2 ~ dgamma(0.01,0.01)
  median1 <- log(2)/theta1
  median2 <- log(2)/theta2
}
```

The model was saved in the file `leukemia.txt` which will be required in the next section. Ultimately, the R output that we will get from the code in the next section will consist of vectors of 50,000 posterior samples for each of `theta1`, `theta2`, `median1`, and `median2`. We can work with these vectors like we would any R vector, e.g., compute means, standard deviations, and obtain plots.

Readers who want to run the R code as they go along should now skip to Section C.6 and, after running that code, return here.

Figure C.5 presents posterior densities, i.e., smoothed histograms of the simulated posterior iterates, for the median survival time in each group. It was obtained using the following R commands:

```
plot(density(median1),type="l",xlab="",main="",ylab="",
      xlim=c(0,75), ylim=c(0,0.15), lwd=3)
```

```

lines(density(median2),lty=2,lwd=3)
mtext("Median survival",line=3,side=1,cex=1.5)
mtext("Posterior density",line=2.5,side=2,cex=1.5)
text(26,0.14,"AG Negative",lwd=2,cex=1.2)
text(61,0.03,"AG Positive",lwd=2,cex=1.2)

```

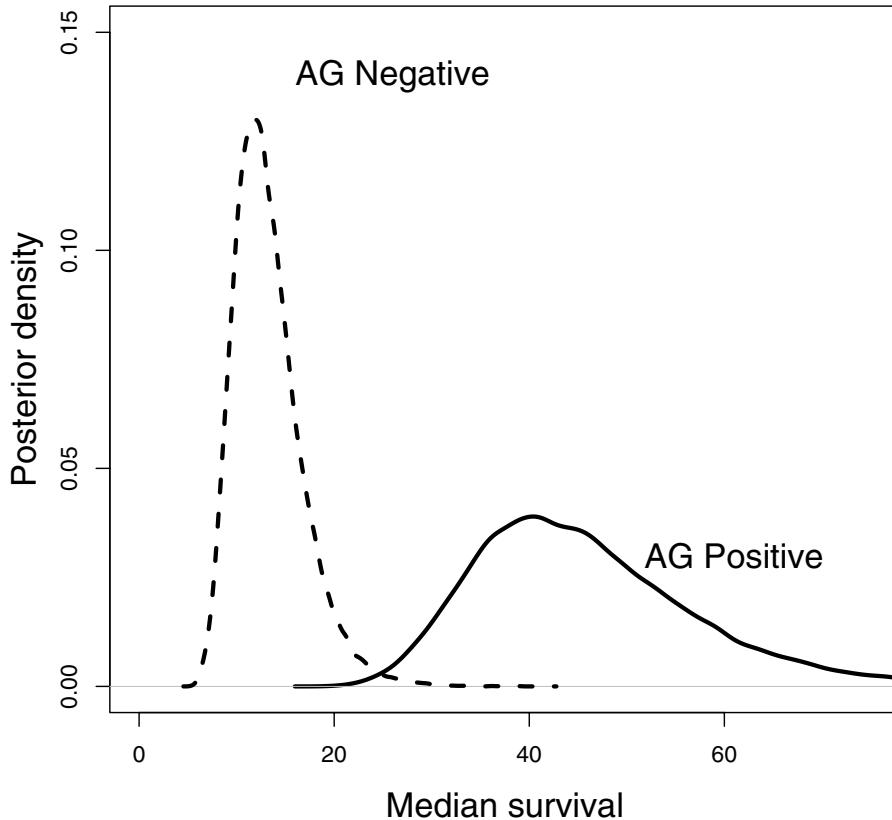


Figure C.5: Posterior densities of median survival times for the leukemia data.

The command `mtext` (marginal text) with `side=1` labels the horizontal axis and `side=2` labels the vertical axis; `line` controls the distance from an axis to the text. The `text` commands place text within the graphic itself where the numbers given are the  $(x,y)$  coordinates for placement of the characters. The argument `cex` stands for character expansion. It controls the size of the displayed text for the axis labels and for the text within the graphic.

Now, back to the task of estimating survival functions. Let's focus on estimating  $S_1(t)$ . The same Monte Carlo approach can be used to estimate  $S_2(t)$ . For each value  $\theta_1^k$  sampled from the posterior  $p(\theta_1|y)$ , we compute  $S_1^k(t) = e^{-\theta_1^k t}$ . We cannot evaluate  $S_1^k(t)$  for all  $t > 0$ , so we chop up the positive real line into a fine grid and evaluate  $S_1$  at each grid point. We might use the grid  $t = 0, 0.01, 0.02, \dots, T_*$ , where  $T_*$  is just past the largest observed time in the data. In the leukemia analysis, we used  $t = 0, 1, 2, \dots, 170$  since the biggest observed time is 156, and steps of 1 unit give sufficient detail across the range 0 to 170. We estimate  $S_1(t)$  at each of these values of  $t$  and then interpolate (connect the dots) to get an estimate of the entire curve. The posterior mean survival

function is approximated as

$$E\{S_1(t)|\theta_1\} \doteq \frac{1}{m} \sum_{k=1}^m e^{-\theta_1^k t}.$$

A pointwise 95% posterior band is obtained by plotting the 2.5 and 97.5 percentiles at each grid point and interpolating.

We have 50,000 simulated values from  $p(\theta_1|y)$ . For illustration, the first 20 sampled survival curves for the AG positive group are plotted in the left panel of Figure (R code provided below).

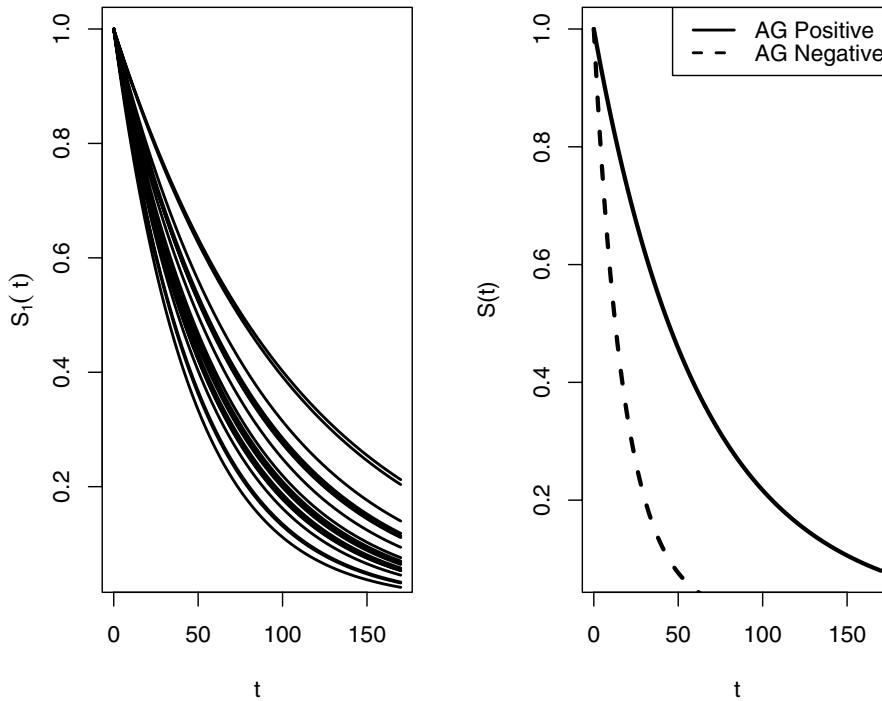


Figure C.6: 20 simulated values of  $S_1(t)$  for the AG positive group (left panel), and the posterior means based on 50,000 simulated survival functions (right panel).

```
time <- seq(0,170,1)
plot(time,exp(-theta1[1]*time),type="l",
     xlab="t",ylab="",cex=2,lwd=2)
mtext(expression(S[1](t)),line=2.5,side=2)
for(i in 2:20){
  lines(time,exp(-theta1[i]*time),lty=1,cex=2,lwd=2)
}
```

To create an axis label that contains a subscript, we have used the `expression` command. Superscripts and Greek letters can also be plotted using this command. For example, when used in the `text` or `mtext` commands, `expression(mu)` will produce  $\mu$  and `expression(theta^1)` produces  $\theta^1$ .

A plot of the mean of the 50,000 posterior iterates of  $S_1(t)$  at  $t = 0, 1, 2, \dots, 170$  is in the right panel of Figure . The same procedure was used to estimate  $S_2(t)$ . The R code that was used to generate the right panel of Figure follows.

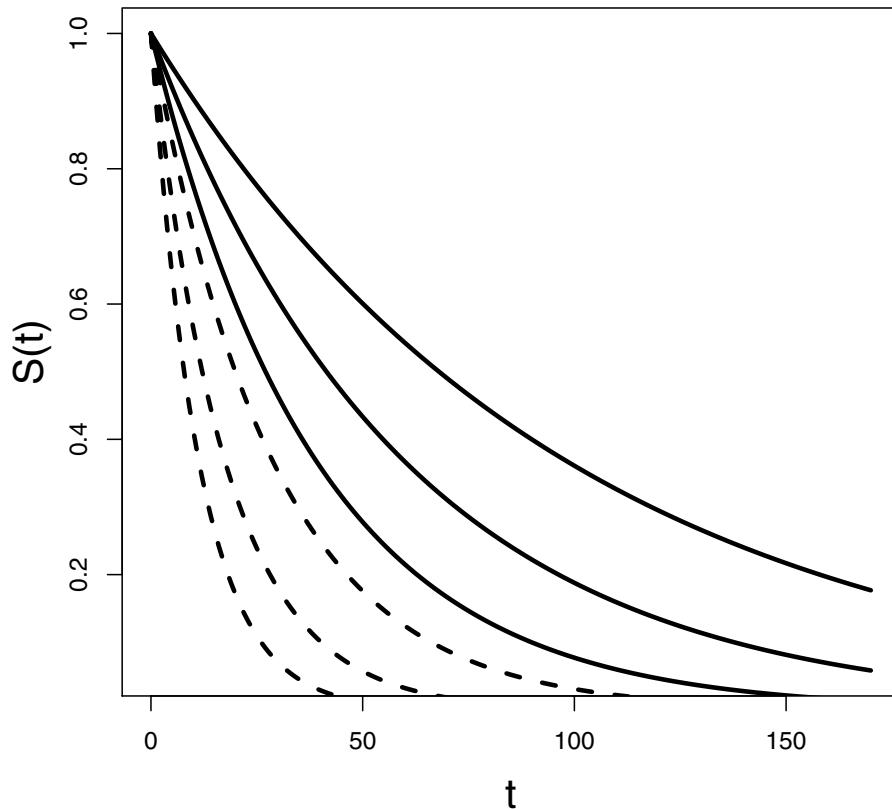


Figure C.7: Posterior median and 95% pointwise probability bands for the AG positive (solid lines) and negative (dashed lines) groups.

```

survcurves1 <- exp(-theta1%*%t(time))
survmean1 <- apply(survcurves1,2,mean)
survcurves2 <- exp(-theta2%*%t(time))
survmean2 <- apply(survcurves2,2,mean)
plot(time,survmean1,type="l",xlab="t",ylab="S(t)",lwd=3)
lines(time, survmean2,lty=2,lwd=3)
legend("topright",c("AG Positive","AG Negative"),lty=1:2,lwd=2)

```

Note that `theta1%*%t(time)` creates a 50,000 by 171 matrix. To place the two graphs side-by-side in Figure , the two portions of R code presented above were preceded by `par(mfrow=c(1, 2))` and then run in tandem.

We can also compute the 2.5, 50, and 97.5 percentiles of the 50,000 iterates at each grid point to obtain approximations to the posterior medians and pointwise probability bands for the two survival functions. In general, graphs that contain 6 survival curves can be a bit cluttered, especially close to  $t = 0$ . However, for the leukemia data the groups are well separated (Figure ). The R code used to create this figure is presented below.

```

surv1 <- apply(survcurves1,2,quantile,c(0.025,0.50,0.975))
surv2 <- apply(survcurves2,2,quantile,c(0.025,0.50,0.975))
plot(time, surv1[2,], type="l", xlab="t", ylab="S(t)", lwd=3)
lines(time, surv1[1,], lty=1, lwd=3)
lines(time, surv1[3,], lty=1, lwd=3)
lines(time, surv2[2,], lty=2, lwd=3)

```

```
lines(time, surv2[1,], lty=2, lwd=3)
lines(time, surv2[3,], lty=2, lwd=3)
```

### C.6 Interface Between R and WinBUGS

We now illustrate the application of R and WinBUGS in tandem. The `R2WinBUGS` library (Sturtz et al., 2005) serves as an R interface to WinBUGS. The main command in this library is `bugs`. After installing and calling the `R2WinBUGS` library, enter `?bugs` to study the various arguments in this function (many of which are explained below).

The following R code was used to generate the posterior samples for the leukemia data. The WinBUGS model was saved as `leukemia.txt` in the `working.directory`. (This path will need to be changed appropriately by the user.)

```
library(R2WinBUGS)
n1 <- 17
n2 <- 16
y <- c(65,156,100,134,16,108,121,4,39,143,56,26,22,1,1,5,65)
x <- c(56,65,17,7,16,22,3,4,2,3,8,4,3,30,4,43)
leukemiadata <- list("n1","n2","y","x")
parameters <- c("theta1","theta2","median1","median2")
inits <- list(list(theta1=1, theta2=1))
leuk.fit <- bugs(leukemiadata, inits, parameters, "leukemia.txt",
                  working.directory="H:\\RApp", n.chains=1, n.iter=50000,
                  n.thin=1, n.burnin=0)
print(leuk.fit)
attach.bugs(leuk.fit)
```

The `print` command reports several posterior summaries for each variable in `parameters`, and the `attach.bugs` command makes all the simulated posterior iterates available for manipulation in R. Recall that the syntax of WinBUGS involves various lists, e.g., for initial parameter values. That syntax is also used here so that the list of initial values for WinBUGS is put into a list that is fed into `bugs`.

### C.7 Writing New R Functions

In addition to its graphical capabilities, another big upside to R is that it enables users to write their own functions. As a simple illustration we create an R function that we named `t.interval` for computing the estimate and a  $t$  confidence interval, oops, the standard improper prior's posterior mean and a probability interval for  $\mu_1 - \mu_2$  from independent, two-group normal data with equal population variances, cf. Section 5.2. The  $t$  confidence interval is given by

$$(\bar{y}_1 - \bar{y}_2) \pm t_{n_1+n_2-2}(1 - \alpha/2) \sqrt{s_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}$$

where  $s_p^2 = \{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2\}/(n_1 + n_2 - 2)$  is the pooled sample variance and the appropriate  $t$  multiplier depends on the confidence level  $1 - \alpha$ . The 3 required inputs to our function `t.interval` are the data vectors from 2 populations (`y1` and `y2`) and the desired confidence level (`conflevel`). The first 2 lines in the function create complete case data vectors by omitting any missing values. The next lines define the sample sizes of the complete cases (`n1` and `n2`), the pooled variance (`sp2`), the difference in sample means (`d`), the  $t$  multiplier (`t`), and the lower and upper endpoints of the confidence interval (`l` and `u`). The function outputs `l`, `d`, and `u`.

```
t.interval=function(y1,y2,conflevel){
  y1c=na.omit(y1)
```

```
y2c=na.omit(y2)
n1=length(y1c)
n2=length(y2c)
sp2=((n1-1)*var(y1c) + (n2-1)*var(y2c))/(n1+n2-2)
d=mean(y1c)-mean(y2c)
t=qt(conflevel+0.5*(1-conflevel),n1+n2-2)
l=d-t*sqrt(sp2*(1/n1+1/n2))
u=d+t*sqrt(sp2*(1/n1+1/n2))
return(c(l,d,u))
}
```

Of course, R comes equipped with a function (called `t.test`) that calculates  $t$  intervals. To illustrate, we simulated 15 observations  $y_i \stackrel{iid}{\sim} N(2, 1)$  and 102 observations  $x_i \stackrel{iid}{\sim} N(0, 1)$  where the  $x$  observations have the R default mean and standard deviation with the last two missing. The following code defines the data and presents the output from `t.interval` and an excerpt of the output from `t.test`.

```
> x <- c(rnorm(100),NA,NA)
> y <- rnorm(15,2,1)

> t.interval(x,y,0.95)
[1] -2.689665 -2.119686 -1.549707

> t.test(x,y, var.equal=TRUE)

95 percent confidence interval:
-2.689665 -1.549707
sample estimates:
mean of x   mean of y
-0.02400025  2.09568550
```



---

## References

---

- Adcock, C.J. (1997). Sample size determination: A review. *The Statistician*, **46**, 261–283.
- Aitchison, J. and Dunsmore, I.R. (1975). *Statistical Prediction Analysis*. Cambridge University Press, Cambridge.
- Albert, J. (2007). *Bayesian Computation with R*. Springer-Verlag, New York.
- Aldrich, J. (2005). Fisher and regression. *Statistical Science*, **20**, 401–417.
- Antoniak, C. (1974). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, **1**, 209–230.
- Arnold, B.C. and Press, S.J. (1989). Compatible conditional distributions. *Journal of the American Statistical Association*, **84**, 152–156.
- Banerjee, S., Carlin, B.P., and Gelfand, A.E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC, Boca Raton, FL.
- Barron, A., Schervish, M.J., and Wasserman, L. (1999). Posterior distributions in nonparametric problems. *Annals of Statistics*, **27**, 536–561.
- Bedrick, E.J., Christensen, R., and Johnson, W. (1996). A new perspective on priors for generalized linear models. *Journal of the American Statistical Association*, **91**, 1450–1460.
- Bedrick, E.J., Christensen, R., and Johnson, W. (1997). Bayesian binomial regression. *The American Statistician*, **51**, 211–218.
- Bedrick, E.J., Christensen, R., and Johnson, W. (2000). Bayesian accelerated failure time analysis with application to veterinary epidemiology. *Statistics in Medicine*, **19**, 221–237.
- Belitz, C., Brezger, A., Kneib, T., and Lang, S. (2009). BayesX – Software for Bayesian inference in structured additive regression models. Version 2.00. Available from <http://www.stat.uni-muenchen.de/~bayesx>.
- Belsley, D.A. (1991). *Collinearity Diagnostics: Collinearity and Weak Data in Regression*. John Wiley and Sons, New York.
- Benjamini, Y. and Hockberg, Y. (1995). Controlling the false discovery rate: A new and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, **57**, 289–300.
- Berger, J.O. (1993). *Statistical Decision Theory and Bayesian Analysis*, Revised Second Edition. Springer-Verlag, New York.
- Berger, J.O. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science*, **18**, 1–32.
- Berger, J.O. and Guglielmi, A. (2001). Bayesian and conditional frequentist testing of a parametric model versus nonparametric alternatives. *Journal of the American Statistical Association*, **96**, 174–184.
- Berger, J.O. and Wolpert, R. (1984). *The Likelihood Principle*. Institute of Mathematical Statistics Monograph Series, Hayward, CA.
- Bernardo, J.M. (1979). Reference posterior distributions for Bayesian inference (with discussion). *Journal of the Royal Statistical Society, Series B*, **41**, 113–147.
- Bernardo, J.M. and Smith, A.F.M. (2000). *Bayesian Theory*. John Wiley and Sons, Chichester, West Sussex.
- Berry, D.A. (1996). *Statistics: A Bayesian Perspective*. Duxbury Press, Belmont, CA.
- Bisgaard, S. and Fuller, H.T. (1996). Reducing variation with two-level factorial experiments. *Quality Engineering*, **8**, 373–377.
- Bissell, A.F. (1972). A negative binomial model with varying element sizes. *Biometrika*, **59**, 435–441.
- Box, G.E.P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society, Series A*, **143**, 383–404.

- Box, G.E.P. and Tiao, G.C. (1973). *Bayesian Inference in Statistical Analysis*. John Wiley and Sons, New York.
- Branscum, A.J., Gardner, I.A., and Johnson, W.O. (2004). Bayesian modeling of animal and herd-level prevalences. *Preventive Veterinary Medicine*, **66**, 101–112.
- Branscum, A.J., Gardner, I.A., and Johnson, W.O. (2005). Estimation of diagnostic test sensitivity and specificity through Bayesian modeling. *Preventive Veterinary Medicine*, **68**, 145–163.
- Branscum, A.J., Perez, A.M., Johnson, W.O., and Thurmond, M.C. (2008). Bayesian spatiotemporal analysis of foot-and-mouth disease data from the Republic of Turkey. *Epidemiology and Infection*, **136**(6), 833–842.
- Brinkman, N.D. (1981). Ethanol fuel – A single-cylinder engine study of efficiency and exhaust emissions. *SAE Transactions*, **90**, 1410–1424.
- Broemeling, L.D. (2007). *Bayesian Biostatistics and Diagnostic Medicine*. Chapman and Hall/CRC, Boca Raton, FL.
- Brooks, S.P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, **7**, 434–455.
- Carey, J.R., Liedo, P., Müller, H.G., Wang, J.L., and Chiou, J.M. (1998). Relationship of age patterns of fecundity to mortality, longevity, and lifetime reproduction in a large cohort of Mediterranean fruit fly females. *Journal of Gerontology Series A: Biological Sciences and Medical Sciences*, **53**, 245–251.
- Carlin, B.P. and Gelfand, A.E. (1991). An iterative Monte Carlo method for nonconjugate Bayesian analysis, *Statistical Computing*, **1**, 119–128.
- Carlin, B.P. and Louis, T.A. (2008). *Bayesian Methods for Data Analysis*, Third Edition. Chapman and Hall/CRC, Boca Raton, FL.
- Casella, G. and George, E.I. (1992). Explaining the Gibbs sampler. *The American Statistician*, **46**, 167–174.
- Chaloner, K. and Brant, R. (1988). A Bayesian approach to outlier detection and residual analysis. *Biometrika*, **75**, 651–660.
- Chen, M.-H., Shao, Q.-M., and Ibrahim, J.G. (2000). *Monte Carlo Methods in Bayesian Computation*. Springer-Verlag, New York.
- Choi, Y.K., Johnson, W.O., Collins, M.T., and Gardner, I.A. (2006). Bayesian estimation of ROC curves in the absence of a gold standard. *Journal of Agricultural, Biological, and Environmental Statistics*, **11**, 210–229.
- Christensen, R. (1996). *Analysis of Variance, Design, and Regression: Applied Statistical Methods*. Chapman and Hall/CRC, Boca Raton, FL.
- Christensen, R. (1997). *Log-Linear Models and Logistic Regression*, Second Edition. Springer-Verlag, New York.
- Christensen, R. (2001a). *Advanced Linear Modeling*, Second Edition. Springer-Verlag, New York.
- Christensen, R. (2001b). Letter to the Editor, *Journal of Quality Technology*, **33**, 127.
- Christensen, R. (2002). *Plane Answers to Complex Questions: The Theory of Linear Models*, Third Edition. Springer-Verlag, New York.
- Christensen, R. (2005). Testing Fisher, Neyman, Pearson, and Bayes. *The American Statistician*, **59**, 121–126.
- Christensen, R. (2007). General prediction theory and the role of  $R^2$ . Unpublished manuscript.
- Christensen, R. (2008). Review of “Principles of Statistical Inference” by D. R. Cox. *Journal of the American Statistical Association*, **103**, 1719–1723.
- Christensen, R. (2009). Inconsistent Bayesian estimation. *Bayesian Analysis*, **4**, 759–762.
- Christensen, R., Hanson, T., and Jara, A. (2008). Parametric nonparametric statistics: An introduction to mixtures of finite Polya trees. *The American Statistician*, **62**, 296–306.
- Christensen, R. and Huffman, M.D. (1985). Bayesian point estimation using the predictive distribution. *The American Statistician*, **39**, 319–321.
- Clyde, M. and George, E.I. (2004). Model uncertainty. *Statistical Science*, **19**, 81–94.
- Colditz, G.A., Stampfer, M.J., Willett, W.C., Hennekens, C.H., Rosner, B., and Speizer, F.E. (1990). Prospective study of estrogen replacement therapy and risk of breast cancer in postmenopausal women. *Journal of the American Medical Association*, **264**, 2648–2653.

- Collett, D. (2003). *Modelling Survival Data in Medical Research*, Second Edition. Chapman and Hall/CRC, Boca Raton, FL.
- Congdon, P. (2001) *Bayesian Statistical Modelling*. John Wiley and Sons, Chichester, West Sussex.
- Cook, R.D. (1977). Detection of influential observations in linear regression. *Technometrics*, **19**, 15–18.
- Cook, R.D. and Weisberg, S. (1999). *Applied Regression Including Computing and Graphics*. John Wiley and Sons, New York.
- Cox, D.R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society, Series B (Methodological)*, **34**, 187–220.
- Cox, D.R. (2006). *Principles of Statistical Inference*. Cambridge University Press, Cambridge.
- Crainiceanu, C.M. and Goldsmith, A.J. (2009). Bayesian functional data analysis using WinBUGS. *Journal of Statistical Software*, **32(11)**, 1–43.
- Crainiceanu, C.M., Ruppert, D., and Wand, M.P. (2004). Bayesian Analysis for Penalized Spline Regression Using WinBUGS. Johns Hopkins University, Dept. of Biostatistics Working Papers. Working Paper 40. <http://www.bepress.com/jhubiostat/paper40>.
- Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press, Princeton.
- Crawley, M.J. (2007). *The R Book*. John Wiley and Sons, Chichester, West Sussex.
- Dalal, S.R., Fowlkes, E.B., and Hoadley, B. (1989) Risk analysis of the space shuttle: Pre-Challenger prediction of failure. *Journal of the American Statistical Association*, **84**, 945–957.
- Dalgaard, P. (2008). *Introductory Statistics with R*, 2nd Edition. Springer-Verlag, New York.
- Davis, C.S., (2002). *Statistical Methods for the Analysis of Repeated Measurements*. Springer-Verlag, New York.
- De Backer, M., De Keyser, P., De Vroey, C. and Lesaffre, E. (1996). A 12-week treatment for dermatophyte toe onychomycosis: terbinafine 250mg/day vs. itraconazole 200mg/day—a double-blind comparative trial. *British Journal of Dermatology*, **134**, 16–17.
- de Boor, C. (1977). Package for calculating with B-splines. *SIAM Journal of Numerical Analysis*, **14**, 441–472.
- de Finetti, B. (1974, 1975). *Theory of Probability*, Vols. 1 and 2. John Wiley and Sons, New York.
- Deal, E.C. Jr., McFadden, E.R. Jr., Ingram, R.H. Jr., Strauss, R.H., and Jaeger, J.J. (1979). Role of respiratory heat exchange in production of exercise-induced asthma. *Journal of Applied Physiology*, **46**, 467–475.
- DeGroot, M.H. (1970). *Optimal Statistical Decisions*. McGraw-Hill, New York.
- Dellaportas, P. and Smith, A.F.M. (1993). Bayesian inference for generalized linear models and proportional hazards via Gibbs sampling. *Applied Statistics*, **42**, 443–459.
- Denison, D.G.T., Holmes, C.C., Mallick, B.K., and Smith, A.F.M. (2002). *Bayesian Methods for Nonlinear Classification and Regression*. John Wiley and Sons, Chichester, West Sussex.
- Eilers, P.H.C. and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, **11**, 89–121.
- Engøe, C., Georgiadis, M.P., and Johnson, W.O. (2000). Evaluation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease status is unknown. *Preventive Veterinary Medicine*, **45**, 61–81.
- Escobar, M.D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, **90**, 577–588.
- Everitt, B.S. and Hothorn, T. (2006). *A Handbook of Statistical Analyses Using R*. Chapman and Hall/CRC, Boca Raton, FL.
- Feigl, P. and Zelen, M. (1965). Estimation of exponential survival probabilities with concomitant information. *Biometrics*, **21**, 826–838.
- Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, **1**, 209–230.
- Ferguson, T.S. (1974). Prior distributions on spaces of probability measures. *Annals of Statistics* **2**, 615–629.
- Fisher, R.A. (1956). *Statistical Methods and Scientific Inference*, Third Edition, 1973. Hafner Press, New York.
- Gamerman, D. and Lopes, H.F. (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*, Second Edition. Chapman and Hall/CRC, Boca Raton, FL.

- Gastwirth, J.L. (1988). *Statistical Reasoning in Law and Public Policy*, Vol. 1; *Statistical Concepts and Issues of Fairness*. Vol. 2. *Tort Law, Evidence and Health*. Academic Press, Orlando, FL.
- Geisser, S. (1971). The inferential use of predictive distributions. In *Foundations of Statistical Inference*, edited by V.P. Godambe and D.A. Sprott. Holt, Reinhart, and Winston, Toronto; 456–469.
- Geisser, S. (1984). On prior distributions for binary trials. *The American Statistician*, **38**, 244–247.
- Geisser, S. (1992). On the curtailment of sampling. *Canadian Journal of Statistics*, **20**, 297–309.
- Geisser, S. (1993). *Predictive Inference: An Introduction*. Chapman and Hall, New York.
- Geisser, S. and Eddy, W.F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, **74**, 153–160.
- Gelfand, A.E. and Dey, D.K. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society, Series B*, **56**, 501–514.
- 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. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, **3**, 515–533.
- Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2004). *Bayesian Data Analysis*, Second Edition. Chapman and Hall/CRC, Boca Raton, FL.
- Gelman, A., Meng, X.-L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, **6**, 733–807.
- Gelman, A. and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457–472.
- Georgiadis, M.P., Gardner, I.A., and Hedrick, R.P. (1998). Field evaluation of sensitivity and specificity of a polymerase chain reaction (PCR) for detection of *N. salmonis* in rainbow trout. *Journal of Aquatic Animal Health*, **10**, 372–380.
- Gilks, W.R., Richardson, S., and Spiegelhalter, D., eds. (1996). *Markov Chain Monte Carlo in Practice*. Chapman and Hall, New York.
- Gilks, W.R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, **41**, 337–348.
- Green, P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.
- Grieve, A.P. (1988). A Bayesian approach to the analysis of LD<sub>50</sub> experiments. In *Bayesian Statistics 3*, edited by J.M. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith. Oxford University Press, Oxford; 617–630.
- Gustafson, P. (2005). On model expansion, model contraction, identifiability and prior information: Two illustrative scenarios involving mismeasured variables, with discussion. *Statistical Science*, **20**, 111–120.
- Haar, A. (1910). Zur theorie der orthogonalen funktionensysteme. *Mathematische Annalen*, **69**, 331–371.
- Hamburg, B.A., Kraemer, H.C., and Jahnke, W.A. (1975). Hierarchy of drug use in adolescence behavior and attitudinal correlates of substantial drug use. *American Journal of Psychiatry*, **132**, 1155–1163.
- Hanson, T. (2006). Inference for mixtures of finite Polya tree models. *Journal of the American Statistical Association*, **101**, 1548–1565.
- Hanson, T.E., Johnson, W.O., and Gardner, I.A. (2003). Hierarchical models for estimating herd prevalence and test accuracy in the absence of a gold standard. *Journal of Agricultural, Biological, and Environmental Statistics*, **8**, 223–239.
- Hanson, T.E., Johnson, W.O., and Laud, P.W. (2009). A unified approach to semiparametric inference for survival models with step process covariates. *Canadian Journal of Statistics*, **37**, 60–79.
- Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.
- Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their application. *Biometrika*, **57**, 97–109.

- Hennerfeind, A., Brezger, A., and Fahrmeir, L. (2006). Geoadditive survival models. *Journal of the American Statistical Association*, **101**, 1065–1075.
- Hicks, C.R. (1956). Fundamentals of analysis of variance. Part I – The analysis of variance (ANOVA) model. *Industrial Quality Control*, **13**(2), 17–20.
- Holmes, C.C. and Mallick, B.K. (2001). Bayesian regression with multivariate linear splines. *Journal of the Royal Statistical Society, Series B*, **63**, 3–17.
- Hui, S.L. and Walter, S.D. (1980). Estimating the error rates of diagnostic tests. *Biometrics*, **36**, 167–171.
- Ibrahim, J.G., Chen, M-H., and Sinha, D. (2001). *Bayesian Survival Analysis*. Springer-Verlag, New York.
- Ibrahim, J.G. and Laud, P.W. (1991). On Bayesian analysis of generalized linear models using Jeffreys's prior. *Journal of the American Statistical Association*, **86**, 981–986.
- Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, **5**, 299–314.
- Ishwaran H. and Zarepour M. (2002). Exact and approximate sum-representations for the Dirichlet process. *Canadian Journal of Statistics*, **30**, 269–283.
- Jara, A. (2007). Applied Bayesian non- and semi-parametric inference using DPpackage. *Rnews*, **7**(3), 17–26.
- Jara, A., Hanson T., and Lesaffre, E. (2009). Robustifying generalized linear mixed models using a new class of mixtures of multivariate Polya trees. *Journal of Computational and Graphical Statistics*, **18**, 838–860.
- Jeffreys, H. (1961). *Theory of Probability*, Third Edition. Oxford University Press, London.
- Johnson, W. (1996). Predictive influence in the lognormal survival model. In *Prediction and Modelling in Statistics and Econometrics: Essays in Honor of Seymour Geisser*, edited by J. Lee, W. Johnson, and A. Zellner. Elsevier, Amsterdam; 104–121.
- Johnson, W.O., Gastwirth, J.L., and Pearson, L.M. (2001). Screening without a gold standard: The Hui-Walter paradigm revisited. *American Journal of Epidemiology*, **153**, 921–924.
- Johnson, W. and Geisser, S. (1983). A predictive view of the detection and characterization of influential observations in regression. *Journal of the American Statistical Association*, **78**, 137–144.
- Johnson, W. and Geisser, S. (1985). Estimative influence measures for the multivariate general linear model. *Journal of Statistical Planning and Inference*, **11**, 33–56.
- Johnson, W.O. and Hanson, T.E. (2005). Comment on ‘On model expansion, model contraction, identifiability and prior information: Two illustrative scenarios involving mismeasured variables.’ *Statistical Science*, **20**, 111–140.
- Johnson, W.O., Su, C-L., Gardner, I.A., and Christensen, R. (2004). Sample size calculations for surveys to substantiate freedom of populations from infectious agents. *Biometrics*, **60**, 165–171.
- Jones, G., Johnson, W.O., Hanson, T.E., and Christensen, R. (2010). Identifiability of models for multiple diagnostic testing in the absence of a gold standard. *Biometrics*, to appear.
- Joseph, L., Gyorkos, T.W., and Coupal, L. (1995). Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *American Journal of Epidemiology*, **141**, 263–272.
- Kadane, J.B and Lazar, N.A. (2004). Methods and criteria for model selection. *Journal of the American Statistical Association*, **99**, 279–290.
- Kalbfleisch, J.D. (1978). Non-parametric Bayesian analysis of survival time data. *Journal of the Royal Statistical Society, Series B*, **40**, 214–221.
- Kalbfleisch, J.D. and Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. John Wiley and Sons, New York.
- Kardaun, O. (1983). Statistical analysis of male larynx-cancer patients - A case study. *Statistica Nederlandica*, **37**, 103–126.
- Kass, R., Tierney, L., and Kadane, J. (1988). Asymptotics in Bayesian computation. In *Bayesian Statistics 3*, edited by J.M. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith. Oxford University Press, Oxford; 261–278.
- Klein, J. and Moeschberger, M.L. (2003). *Survival Analysis Techniques for Censored and Truncated Data*, Second Edition. Springer-Verlag, New York.

- Krause, A. and Olson, M. (2005). *The Basics of S-PLUS*, Fourth Edition. Springer-Verlag, New York.
- Lang, S. and Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics*, **13**, 183–212.
- Lavine, M. (1991). Problems in extrapolation illustrated with space shuttle o-ring data (with discussion). *Journal of the American Statistical Association*, **86**, 919–921.
- Lavine, M. (1992). Some aspects of Polya tree distributions for statistical modeling. *Annals of Statistics*, **20**, 1222–1235.
- Lavine, M. (1994). More aspects of Polya tree distributions for statistical modeling. *Annals of Statistics*, **22**, 1161–1176.
- Lehmann, E.L. (1986). *Testing Statistical Hypotheses*, Second Edition. John Wiley and Sons, New York.
- Leonard, T. (1982). Comment on ‘A simple predictive density function,’ by Lejeune and Faulkenberry. *Journal of the American Statistical Association*, **77**, 657–658.
- Lesaffre, E. and Spiessens, B. (2001). On the effect of the number of quadrature points in a logistic random-effects model: An example. *Applied Statistics*, **50**, 325–335.
- Li, L., Hu, B., and Greene, T. (2009). A semiparametric joint model for longitudinal and survival data with application to hemodialysis study. *Biometrics*, **65**, 737–745.
- Lindley, D.V. (1971). *Bayesian Statistics: A Review*. SIAM, Philadelphia.
- Lo, A.Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *Annals of Statistics*, **12**, 351–357.
- Lunn, D., Spiegelhalter, D., Thomas, A., and Best, N. (2009). The BUGS project: Evolution, critique, and future directions. *Statistics in Medicine*, **28**, 3049–3067.
- Lunn, D.J., Thomas, A., Best, N.G., and Spiegelhalter, D.J. (2000). WinBUGS – A Bayesian modelling framework: Concepts, structure and extensibility. *Statistics and Computing*, **10**, 321–333.
- McMillan, G.P. (2001). Ache Residential Grouping and Social Foraging. Unpublished PhD Dissertation, Dept. of Anthropology, The University of New Mexico.
- Mallick, B.K., Denison, D.G.T., and Smith, A.F.M. (1999). Bayesian survival analysis using a MARS model. *Biometrics*, **55**, 1071–1077.
- Marin, J-M. and Robert, C.P. (2007). *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*. Springer-Verlag, New York.
- Marshal, K. (2003). Personal communication.
- Martz, H. F. and Zimmer, W. J. (1992). The risk of catastrophic failure of the solid rocket boosters on the space shuttle. *The American Statistician*, **46**, 42–47.
- MathSoft, Inc. (1999). *S-Plus 5 for UNIX Guide to Statistics*, Data Analysis Products Division, MathSoft, Seattle.
- Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087–1092.
- Mosteller, F. and Tukey, J. W. (1977). *Data Analysis and Regression*. Addison-Wesley, Reading, MA.
- Müller, P. and Vidakovic, B. (1999). *Bayesian Inference in Wavelet-based Models*. Springer-Verlag, New York.
- Naylor, J.C. and Smith, A.F.M. (1982). Applications of a method for the efficient computation of posterior distributions. *Applied Statistics*, **31**, 214–225.
- Paddock, S.M. (1999). Randomized Polya Trees: Bayesian Nonparametrics for Multivariate Data Analysis. Unpublished doctoral thesis, Institute of Statistics and Decision Sciences, Duke University.
- Paré, J., Thurmund, M., and Hietala, S. (1997). *Neospora caninum* antibodies in cows during pregnancy as a predictor of congenital infection and abortion. *Journal of Parasitology*, **83**, 82–87.
- Pepe, M.S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, New York.
- Perry, L., Van Dyke, R., and Theye, R. (1974). Sympathoadrenal and hemodynamic effects of isoflurane, halothane, and cyclopropane in dogs. *Anesthesiology*, **40**, 465–470.

- Potthoff, R.F. and Roy, S.N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, **51**, 313–326.
- Press, S.J. and Tanur, J.M. (2001). *The Subjectivity of Scientists and the Bayesian Approach*. Wiley, New York.
- R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.
- Racine-Poon, A., Grieve, A.P., Flühler, H., and Smith, A.F.M. (1986). Bayesian methods in practice: Experiences in the pharmaceutical industry (with discussion). *Applied Statistics*, **35**, 93–150.
- Raiffa, H. and Schlaifer, R. (1961). *Applied Statistical Decision Theory*. Division of Research, Graduate School of Business Administration, Harvard University, Boston.
- Ratkowsky, D (1983). *Nonlinear Regression Modeling*. Marcel Dekker, New York.
- Rice, J.A. (1995). *Mathematical Statistics and Data Analysis*, Second Edition. Duxbury Press, Belmont, CA.
- Robert, C.P. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*, Second Edition. Springer: New York.
- Robert, C.P. and Casella, G. (2004). *Monte Carlo Statistical Methods*, Second Edition. Springer, New York.
- Roberts, G.O. (1996). Markov chain concepts related to sampling algorithms. In *Markov Chain Monte Carlo in Practice*, edited by W.R. Gilks, S. Richardson, and D.J. Spiegelhalter. Chapman and Hall/CRC, Boca Raton, FL; 45–58.
- Roeder, L. (1990). Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association*, **85**, 617–624.
- Rosner, B. (2006). *Fundamentals of Biostatistics*, Sixth Edition. Duxbury Press, Belmont, CA.
- Ross, S. (2006). *A First Course in Probability*, Seventh Edition. Prentice Hall, Englewood Cliffs, NJ.
- Royston, P. (2001). Flexible parametric alternatives to the Cox model, and more. *The Stata Journal*, **1**, 1–28.
- Rubin, D.B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, **12**, 1151–1172.
- Ryan, T.P. and Woodall, W.H. (2005). The most cited statistical papers. *Journal of Applied Statistics*, **32**, 461–474.
- Savage, L.J. (1954). *The Foundations of Statistics*. John Wiley and Sons, New York.
- Schafer, D.W. (1987). Measurement error diagnostics and the sex discrimination problem. *Journal of Business and Economic Statistics*, **5**, 529–537.
- Scheffé, H. (1959). *The Analysis of Variance*. John Wiley and Sons, New York.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.
- Sedmak, D.D., Meineke, T.A., Knechtges, D.S., and Anderson, J. (1989). Prognostic significance of cytokeratin-positive breast cancer metastases. *Modern Pathology*, **2**, 516–520.
- Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, **4**, 639–650.
- Shumway, R.H. (1988). *Applied Statistical Time Series Analysis*. Prentice Hall, Englewood Cliffs, NJ.
- Shumway, R.H. and Stoffer, D.S. (2006). *Time Series Analysis and Its Applications: With R Examples*, Second Edition. Springer-Verlag, New York.
- Smith, A.F.M. and Gelfand, A.E. (1992). Bayesian statistics without tears: A sampling-resampling perspective. *The American Statistician*, **46**, 84–88.
- Smith, A.F.M., Skene, A.M., Shaw, J.E.H., Naylor, J.C., and Dransfield, M. (1985). The implementation of the Bayesian paradigm. *Communications in Statistics – Theory and Methods*, **14**, 1079–1102.
- Smith, M. and Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *Journal of Econometrics*, **75**, 317–343.
- Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B*, **64**, 583–639.
- Sturtz, S., Ligges, U., and Gelman, A. (2005). R2WinBUGS: A package for running WinBUGS from R. *Journal of Statistical Software*, **12**, 1–16.

- Thomas, M.V., Branscum, A., Miller, C.S., Ebersole, J., Al-Sabbagh, M., and Schuster, J.L. (2009). Within-subject variability in repeated measures of salivary analytes in healthy adults. *Journal of Periodontology*, **80**, 1146–1153.
- Tierney, L. (1994). Markov Chains for Exploring Posterior Distributions. *Annals of Statistics*, **22**, 1701–1728.
- Tierney, L. and Kadane, J. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, **81**, 82–86.
- Tsutakawa, R.K. (1975). Bayesian inference for bioassay. Technical Report No. 52. Department of Statistics, Univ. of Missouri – Columbia.
- Tsutakawa, R.K. and Lin, H.Y. (1986). Bayesian estimation of item response curves. *Psychometrika*, **51**, 251–267.
- Turnbull, B.W. and Weiss, L.A. (1978). A likelihood ratio statistic for testing goodness of fit with randomly censored data. *Biometrics*, **34**, 367–375.
- Tuyl, F., Gerlach, R., and Mengersen, K. (2008a). A comparison of Bayes-Laplace, Jeffreys, and other priors: The case of zero events. *The American Statistician*, **62**, 40–44.
- Tuyl, F., Gerlach, R., and Mengersen, K. (2008b). Posterior predictive arguments in favor of the Bayes-Laplace prior as the consensus prior for binomial and multinomial parameters. *Bayesian Analysis*, **4**, 151–158.
- Utts, J. (1991). Replication and meta-analysis in parapsychology (with discussion). *Statistical Science*, **6**, 363–403.
- Venables, W.N. and Ripley, B.D. (2002). *Modern Applied Statistics with S*, Fourth Edition. Springer-Verlag, New York.
- Vidakovic, B. (1998). Wavelet-based nonparametric Bayes methods. In *Practical Nonparametric and Semiparametric Bayesian Statistics*, edited by D. Dey, P. Müller, and D. Sinha. Springer-Verlag, New York; 133–155.
- Wang, F. and Gelfand, A.E. (2002). A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. *Statistical Science*, **17**, 193–208.
- Watkins, D., Bergman, A., and Horton, R. (1994). Optimization of tool life on the shop floor using design of experiments. *Quality Engineering*, **6**, 609–620.
- Weiner, D.A., Ryan, T.J., McCabe, C.H., Kennedy, J.W., Schloss, M., Tristani, F., Chaitman, B.R., and Fisher, L.D. (1979). Correlations among history of angina, ST-segment response and prevalence of coronary-artery disease in the Coronary Artery Surgery Study (CASS). *New England Journal of Medicine*, **301**, 230–235.
- West, M. (1985). Generalized linear models: Scale parameters, outlier accommodation and prior distributions. In *Bayesian Statistics 2*, edited by J.M. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith. North-Holland, Amsterdam; 531–557.
- Wieand, S., Gail, M.H., James, B.R., and James, K.L. (1989). A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika*, **76**, 585–592.
- Wilks, S.S. (1962). *Mathematical Statistics*. John Wiley and Sons, New York.
- Wolfinger, R.D. (1996). Heterogeneous variance: Covariance structures for repeated measures. *Journal of Agricultural, Biological, and Environmental Statistics*, **1**, 205–230.
- Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. John Wiley and Sons, New York.
- Zellner, A. and Rossi, P.E. (1984). Bayesian analysis of dichotomous quantal response models. *Journal of Econometrics*, **25**, 365–393.