Non-additive effects in Logistic Regression

by

Kazi Mahbubur Rashid Azad

M.Sc., University of Dhaka, 1992

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

THE FACULTY OF GRADUATE STUDIES

(Department of Statistics)

We accept this thesis as conforming
to the required standard

The University of British Columbia

March 2003

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of STATISTICS

The University of British Columbia
Vancouver, Canada

Date March 06, 2003

Abstract

Logistic regression is commonly used in epidemiology to model the relationship between risk factors and presence/absence of a disease. Usually it is difficult to look for interaction structure (many possible pairwise interactions, for instance) to include in the model. So a model which is additive on the logit scale is fitted. If the number of risk factors is relatively large such an additive relationship may not make good sense. A new logistic regression model is proposed to incorporate non-additive interaction effects. In some scenarios this model might better reflect the relationship between the response variable and the risk factors. The Bayesian approach is followed to fit the model and a Markov chain Monte Carlo (MCMC) algorithm, known as the hybrid algorithm is used to simulate the parameters. We apply the new model to three examples and interpret the parameter estimates. We compare the predictive performance of the new model with that of the step-wise and the ordinary logistic regression models.

Contents

List of Tables

List of Figures

Acknowledgements

First of all thanks to Almighty and the Department of Statistics to give me the opportunity to study here. I am very grateful to my supervisor Dr. Paul Gustafson for his innovative ideas that he shared with me and many many thanks to him for giving me continuous support regarding theoretical development and implementing them into my thesis. Without his help in programming I could not have finished my thesis. I am thakful to Dr. Lang Wu for his comments about my thesis as the second reader. Many thanks to Lisa Kuramoto for her hints regarding some programming problems and Latex. I am grateful to my wife Nafisa Anawer for her continuous mental support in completing the thesis, though she was not here. I acknowledge my parents' support.

KAZI MAHBUBUR RASHID AZAD

The University of British Columbia
March 2003

To my Parents and Wife.

Chapter 1

Introduction

Regression methods are commonly used to describe the relationship between a response variable and one or more explanatory variables. When the response variable is discrete, taking two (or more) possible values, logistic regression is the standard method of data analysis. In epidemiology we often have to model the relationship between the status (presence/absence) of a disease and one or more risk factors. Logistic regression models are commonly used for describing this type of relationship and determining if there is a significant effect of the risk factors on the disease outcome. Often we have to include interaction effects in the model along with the main effects to predict the response. Usually it is very difficult to look for interaction structure (many possible pairwise interactions, for instance) especially when there are many risk factors to include in the model. Linde and Osius [13] have commented, "within the setting of parametric logistic regression, interactions can be modeled only in a clumsy and limited way". So a model which is additive on the logit scale is fitted. For simplicity assume the risk factors are binary for now. When the number of risk factors are relatively large and if we want to include the pairwise interactions also such an additive relationship may not make good sense. Suppose we have ten risk factors. Further suppose that in fitting the additive model without interactions, each coefficient is estimated to be log(1.5), that is, each risk factor by itself has an odds-ratio of 1.5. The additive model says that someone with all ten risk factors has an odds-ratio of $1.5^{10} \approx 58$ relative to someone without any risk factors. In many applications this would be implausibly large. Thus instead of modeling an additive relationship and trying to pick out pairwise interactions terms, can we fit non-additive models? Are there any other kind

of interactions that better explain the relationship between the outcome of the response variable and the risk factors? The objective of our thesis is to address these questions by considering different functional forms for modeling the regression relationship. We also want to show that non-additive effects might have plausible and reasonable interpretations in real-life situations.

In this chapter we will briefly describe the background materials to understand the logistic regression model, Bayesian approach to parameter estimation, MCMC methods, stepwise logistic regression and cross-validation to find out the predictive performance of different models.

1.1 Logistic Regression

Let Y be a binary response variable indicating presence ($Y = 1$) and absence ($Y = 0$) of a disease. Consider a collection of p explanatory variables or risk factors denoted by the vector $X' = (X_1, X_2, \ldots, X_p)$. In any regression problem the key quantity is the mean value of the response variable, given the value of the independent or explanatory variable. This quantity is called the conditional mean and will be expressed as "$E(Y|X = x)$". Let us denote this mean as $\pi(x)$. For a binary response variable $E(Y|X = x) = 1 \times Pr(Y = 1|X) + 0 \times Pr(Y = 0|X) = Pr(Y = 1|X = x)$. In linear regression we assume that this conditional mean may be expressed as a linear function of X. That is,

$$\begin{aligned} E(Y|X = x) &= \pi(x) \\ &= Pr(Y = 1|X = x) \\ &= \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p, \end{aligned} \tag{1.1}$$

which is called a *linear probability model.* Here β_0 is the intercept and $\beta' = (\beta_1, \beta_2, \ldots, \beta_p)$ are regression coefficients to be estimated. When observations on Y are independent, this model is a *generalized linear model* (GLM) with identity link function. This linear model (1.1) has a major structural defect. With a dichotomous response variable π takes values between 0 and 1, whereas linear functions take values over the entire real line. Model (1.1) predicts $\pi < 0$ and $\pi > 1$ for sufficiently large or small x values. Therefore, there must in fact be a nonlinear relationship between $\pi(x)$ and x.

1.2 Logistic Regression Model

Because of the structural problems with the linear probability model (1.1), a nonlinear or curvilinear relationship between x and $\pi(x)$ is more reasonable. When we expect a monotonic relationship, the S-shaped curve in Figure 1.1 is natural shape for explaining the relationship between a dichotomous response variable and risk factors. A function having this shape is

$$\begin{aligned} \pi(x) &= \frac{\exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)} \\ &= \frac{\exp(\beta_0 + \beta' x)}{1 + \exp(\beta_0 + \beta' x)}, \end{aligned} \tag{1.2}$$

called the *logistic regression* function, where x denotes the vector of predictors.

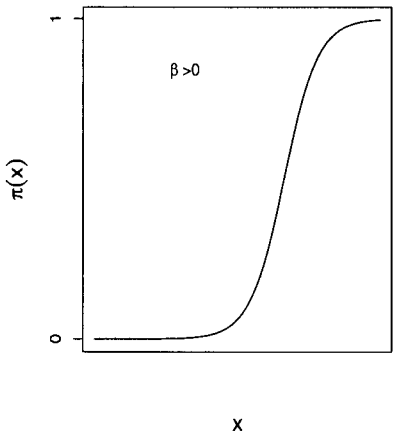

Figure 1.1: Logistic Function

We now have to find the link function that will connect the mismatched quantities and for which the logistic regression model (1.2) is a GLM. For this model the odds of obtaining response $Y = 1$ are

$$\begin{aligned} \frac{\pi(x)}{1 - \pi(x)} &= \exp(\beta_0 + \beta' x) \\ &= e^{\beta_0} e^{\beta_1 x_1} \ldots e^{\beta_p x_p} \end{aligned}$$

This formula provides a basic interpretation for the coefficients. The odds increase multiplicatively by e^{β} for every unit increase in $X = x$. Suppose we are interested in the odds ratio ψ. If we have a single risk factor, say x_1, then ψ for a specific $x_1 = a$ compared to $x_1 = b$ can be computed as:

$$\begin{aligned} \psi &= \frac{\exp(\beta_0 + \beta_1 a)}{\exp(\beta_0 + \beta_1 b)} \\ &= e^{\beta_1(a-b)} \end{aligned}$$

However, in multi-variable situation ψ for a specific risk factor can be computed by keeping all other variables fixed at some arbitrary values, usually at their average values. This is called adjusted odds-ratio.

The log odds has the linear relationship

$$\begin{aligned} \log\left(\frac{\pi(x)}{1-\pi(x)}\right) &= g(x;\beta) \\ &= \beta_0 + \beta' x \end{aligned} \tag{1.3}$$

Thus, the appropriate link is the log odds transformation or the *logit* transformation. The logit, $g(x;\beta)$ is linear in its parameters, may be continuous, and may range from $-\infty$ to $+\infty$, depending on the range of x.

1.3 Model Fitting

1.3.1 Log Likelihood for Binomial Data

Suppose that we have total n number of cases in the study. The responses $y_1, y_2, \ldots, y_n$ are assumed to be the observed values of independent random variables $Y_1, Y_2, \ldots, Y_n$ such that $Y_i, i = 1, 2, \ldots, n$ has the binomial distribution with index m_i, the number of observations in each group and parameter π_i. The y_i's may be success counts or success proportions in each group. For simplicity, we assume $m_i = 1,\ i = 1, 2, \ldots, n$ so that each y_i now represents "success" or "failure" of the outcome. Let $x_i = (x_{i0}, x_{i1}, \ldots, x_{ip})$ denote the pth setting of values of p explanatory variable X's, where $x_{i0} = 1$. The x_i's may be binary or continuous. We express

the logistic regression model (1.3) as

$$\pi(x_i) = \frac{\exp\left(\sum_{j=0}^{p} \beta_j x_{ij}\right)}{\left[1 + \exp\left(\sum_{j=0}^{p} \beta_j x_{ij}\right)\right]} \tag{1.4}$$

Therefore, the likelihood function may be written in the form

$$\begin{aligned} L(\beta) &= \prod_{i=1}^{n} \pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i} \\ &= \left\{\prod_{i=1}^{n}(1-\pi(x_i))\right\}\left\{\prod_{i=1}^{n}\exp\left[\log\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right)^{y_i}\right]\right\} \\ &= \left\{\prod_{i=1}^{n}(1-\pi(x_i))\right\}\exp\left[\sum_{i=1}^{n} y_i \log\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right)\right]. \end{aligned}$$

For model (1.3), the *ith* logit is $\sum_j \beta_j x_{ij}$, so the exponential term in the last expression equals

$$\exp\left[\sum_{i=1}^{n} y_i\left(\sum_{j=0}^{p}\beta_j x_{ij}\right)\right] = \exp\left[\sum_{j=0}^{p}\left(\sum_{i=1}^{n} y_i x_{ij}\right)\beta_j\right].$$

Also, since $[1-\pi(x_i)] = [1+\exp(\sum_j \beta_j x_{ij})]^{-1}$, the log likelihood equals

$$l(\beta) = \sum_{j=0}^{p}\left(\sum_{i=1}^{n} y_i x_{ij}\right)\beta_j - \sum_{i=1}^{n}\log\left[1+\exp\left(\sum_{j=0}^{p}\beta_j x_{ij}\right)\right]. \tag{1.5}$$

The maximum likelihood estimate $\hat{\beta}$ of β satisfies the likelihood equations

$$\frac{\partial l(\beta)}{\partial \beta_j} = 0$$

for $j = 0, 1, \ldots, p$. Since

$$\frac{\partial l(\beta)}{\partial \beta_j} = \sum_i y_i x_{ij} - \sum_i x_{ij}\left[\frac{\exp\left(\sum_j \beta_j x_{ij}\right)}{1+\exp\left(\sum_j \beta_j x_{ij}\right)}\right],$$

the (p+1) likelihood equations are

$$\begin{aligned} \sum_i (y_i - \hat{\pi}_i) &= 0 \\ \text{and} \quad \sum_i (y_i - \hat{\pi}_i) x_{ij} &= 0 \quad \text{for} \quad j = 1, 2, \ldots, p \end{aligned} \tag{1.6}$$

where

$$\hat{\pi}_i = \frac{\exp\left(\sum_{j=0}^{p} \hat{\beta}_j x_{ij}\right)}{\left[1 + \exp\left(\sum_{j=0}^{p} \hat{\beta}_j x_{ij}\right)\right]},$$

denotes the ML estimate of $\pi(x_i)$. Since this system of equations is not linear in β, iterative methods such as Newton-Raphson method are needed to evaluate parameter estimates.

1.4 Bayesian Approach to Model Fitting

In the previous section we described the maximum likelihood estimation(MLE) procedure for estimating the parameters in which we do not incorporate any *prior* information about the parameters, but rather estimate them on the basis of the observed data. Let our parameters of interest be the vector θ, with a precise meaning in the problem under study. In Bayesian data analysis we assume that θ has some probability distribution and we include these information along with the observed data in the estimation process (see, for example, Gelman et. al. [5]).

It is likely that the researcher has some knowledge about θ. Inclusion of this body of knowledge in the analysis is possible and scientifically recommended. Bayesians and frequentists have divergent views in this respect. The latter does not admit this information because it has not been observed and is therefore not subject to empirical verification. The Bayesian approach incorporates this information to the analysis through a density $p(\theta)$ even when this information is not precise. The process of Bayesian data analysis can be divided into the following three steps:

1. Setting up a *full probability model* – a joint probability distribution for all observable and unobservable quantities in a problem.
2. Conditioning on observed data: calculating and interpreting the appropriate *posterior distribution* – the conditional probability distribution of the unobserved quantities of ultimate interest, given the observed data.

3. Evaluating the fit of the model and the implications of the resulting posterior distribution.

Bayesian Inference

Bayesian statistical conclusions about a parameter vector θ, or unobserved data $\tilde{y}$, are made in terms of *probability* statements. These probability statements are conditional on the observed value of y, and we write them as $p(\theta|y)$ or $p(\tilde{y}|y)$. Conditioning also applies to the fixed values of any covariates, x.

Bayes' rule

As outlined in item (1), Bayesian inference contains two ingredients for calculating the posterior density: $p(\theta)$, the *prior distribution* of θ, and $p(y|\theta)$, the *sampling distribution* of the observed data y. The former distribution can also be specified by some constants/parameters just like the distribution of y. Sometimes it is useful to distinguish them from the parameter of interest of θ. These constants are then called hyperparameters, as they are the parameters of the distribution of the parameters. Initially, the hyperparameters are assumed to be known. We call $p(\theta)$ the prior density as it contains the probability distribution of θ *before* the observation of the value of y. The likelihood function of θ is $L(\theta) = p(y|\theta)$. The joint probability distribution of θ and y can then be defined as

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

Simply conditioning on the known value of the data y, using the basic property of conditional probability known as Bayes' rule, yields the *posterior* density:

$$\begin{aligned} p(\theta|y) &= \frac{p(\theta, y)}{p(y)} \\ &= \frac{p(\theta)p(y|\theta)}{p(y)}, \end{aligned} \tag{1.7}$$

where $p(y) = \sum_\theta p(\theta)p(y|\theta)$, and the sum is over all possible values of θ (or $p(y) = \int p(\theta)p(y|\theta)d\theta$ in the case of continuous θ), is called the marginal distribution of y. Since $p(y)$ does not depend on θ, an equivalent expression of (1.7) is the following *unnormalized posterior density*,

$$p(\theta|y) \propto p(\theta)p(y|\theta), \tag{1.8}$$

where the proportionality is as a function of θ for fixed y and $p(y)$ is the normalization constant which may not be evaluated easily.

The concepts of prior and posterior are always relative to the observation considered at a given moment. It is possible that after observing y and obtaining the posterior, a new observation $\tilde{y}$ also related to θ through an eventually different likelihood function becomes available. In this case, the posterior (relative to y) is the prior (relative to $\tilde{y}$) and a new posterior can be obtained by a new application of Bayes' theorem.

A Bayes' estimator of θ is the mean of the posterior distribution of θ, called the posterior expectation, i.e.

$$\begin{aligned} \hat{\theta}_B &= E(\theta|y) \\ &= \int \theta p(\theta|y) d\theta \\ &= \frac{\int \theta p(\theta) p(y|\theta) d\theta}{\int p(\theta) p(y|\theta) d\theta}. \end{aligned} \tag{1.9}$$

Therefore, the obvious distinction between the MLE and Bayesian approach is that in the MLE procedure we estimate the parameters by maximizing the likelihood function whereas in the Bayesian approach we obtain estimates by computing posterior mean of the parameters.

Prediction

To make predictive inferences about unknown observable, $\tilde{y}$, the *posterior predictive distribution* of $\tilde{y}$ is given by:

$$\begin{aligned} p(\tilde{y}|y) &= \int p(\tilde{y}, \theta|y) d\theta \\ &= \int p(\tilde{y}|\theta, y) p(\theta|y) d\theta \\ &= \int p(\tilde{y}|\theta) p(\theta|y) d\theta. \end{aligned} \tag{1.10}$$

From the second and third lines of the equation we see that the posterior predictive distribution is an average of conditional predictions over the posterior distribution of θ.

1.5 MCMC methods for parameter estimation

If θ has high dimension and the model is complex so that we can not get any closed or nice mathematical form for conditional distributions for θ, then it's very difficult to estimate the

parameters. *Markov Chain Monte Carlo* (MCMC) techniques provide an answer to the difficult problem of simulation from the high- dimensional distribution of the unknown quantities that appears in complex models (ref. Gamerman [4], Gilks et. al. [7]). MCMC is essentially Monte Carlo integration using Markov Chains. We need to integrate over the posterior distribution of model parameters given the data. *Monte Carlo* integration draws samples from the required distribution until it approaches equilibrium, known as the limiting distribution, and then forms sample averages to approximate expectations. So our limiting distribution is usually the posterior distribution. *Markov Chain* Monte Carlo draws these samples by running a cleverly constructed Markov Chain for a long time.

The integrations in (1.9) have until recently been the source of most of the practical difficulties in Bayesian inference, especially in high dimensions. In most applications, analytic evaluation of $E(\theta|y)$ is impossible. The best alternative way of evaluation is the MCMC.

1.5.1 Monte Carlo Integration

Let X be a vector of k random variables, with distribution $\pi(.)$, where X consists of model parameters and missing data. Our task is to evaluate the expectation

$$E[f(X)] = \frac{\int f(x)\pi(x)dx}{\int \pi(x)dx}$$

for some function of interest $f(.)$. Monte Carlo integration evaluates $E[f(X)]$ by drawing samples $\{X_t, t = 1, 2, \ldots, n\}$ from $\pi(.)$ and then approximating

$$E[f(X)] \approx \frac{1}{n}\sum_{t=1}^{n} f(X_t).$$

So the population mean of $f(X)$ is estimated by a sample mean. When the samples $\{X_t\}$ are independent, the laws of large numbers ensure that the approximation can be made as accurate as desired by increasing the sample size n.

In general, drawing samples $\{X_t\}$ independently from $\pi(.)$ is difficult. However, the $\{X_t\}$ need not necessarily be independent. They can be generated by any process which draws samples throughout the support of $\pi(.)$ in the correct proportions. One way of doing this is through a Markov Chain having $\pi(.)$ as its stationary distribution. This is then *Markov chain* Monte Carlo.

1.5.2 Markov Chains

A Markov chain is a special type of stochastic processes in which the future state of the process depends only on the present state and is independent of the previous states. Suppose we generate a sequence of random variables, $\{X_0, X_1, X_2, \ldots\}$, such that at each time $t \geq 0$, the next state X_{t+1} is sampled from a distribution $P(X_{t+1}|X_t)$ which depends only on the current state of the chain, X_t. That is, given X_t, the next state X_{t+1} does not depend further on the history of the chain $\{X_0, X_1, , \ldots, X_{t-1}\}$. This sequence is called a *Markov chain*, and $P(.|.)$ is called the *transition kernel* of the chain. we will assume that the chain is time-homogeneous, i.e. $P(.|.)$ does not depend on t.

Let the distribution of X_t given X_0 be denoted by $P^{(t)}(X_t|X_0)$. Subject to regularity conditions, the chain will gradually 'forget' its initial state X_0, and $P^{(t)}(.|X_0)$ will eventually converge to a unique *stationary* distribution, which does not depend on t or X_0. Thus as t increases, the sampled points $\{X_t\}$ will look increasingly like dependent samples from $\pi(.)$. We can now use the output from the Markov chain to estimate $E[f(X)]$, where X has distribution $\pi(.)$. The estimator

$$\bar{t}_n = \frac{1}{n}\sum_{t=1}^{n} f(X_t). \tag{1.11}$$

is called an *ergodic average.* If the chain is ergodic and $E_\pi[f(X)] < \infty$ for the unique limiting distribution π then

$$\bar{t}_n \to E_\pi[f(X)] \text{ as } n \to \infty, \text{ with probability } 1$$

This result is a Markov chain equivalent of the law of large numbers. It states that the averages of chain values also provide strongly consistent estimates of parameters of the limiting distribution π despite their dependence.

1.5.3 The Metropolis-Hastings Algorithm

Markov chains can be constructed in several ways. We will describe Markov chains under the class of Metropolis-Hastings algorithm. This name comes from papers by Metropolis et al. [15] and Hastings [11]. For the Metropolis-Hastings algorithm, at each time t, the next state X_{t+1} is chosen by first sampling a *candidate* point Y from a *proposal* transition $q(.|X_t)$. The proposal

distribution may depend on the current point X_t. The candidate point Y is then *accepted* with probability $\alpha(X_t, Y)$ where

$$\alpha(X, Y) = min\left(1, \frac{\pi(Y)q(X|Y)}{\pi(X)q(Y|X)}\right).$$

If the candidate point is accepted, the next state becomes $X_{t+1} = Y$. If the candidate is rejected, the chain does not move, i.e. $X_{t+1} = X_t$.

Consider a distribution π from which a sample must be drawn via Markov chains. This task will only make sense if the non-iterative generation of π is very complicated or expensive. In this case, a transition kernel $P(X_{t+1}|X_t)$ must be constructed in a way such that π is the equilibrium distribution of the chain. Consider reversible chains where the kernel P satisfies

$$\pi(X_t)P(X_{t+1}|X_t) = \pi(X_{t+1})P(X_t|X_{t+1}).$$

The kernel $P(X_{t+1}|X_t)$ consists of two elements: an arbitrary transition kernel $q(X_{t+1}|X_t)$ and the probability $\alpha(X_t, X_{t+1})$ such that

$$P(X_{t+1}|X_t) = q(X_{t+1}|X_t)\alpha(X_t, X_{t+1}), \quad \text{if} \quad X_t \neq X_{t+1}.$$

So the transition kernel defines a density $P(.|X_t)$ for every possible value of the parameter different from X_t. Consequently, there is a positive probability left for the chain to remain at X_t given by

$$P(X_t|X_t) = 1 - \int q(X_{t+1}|X_t)\alpha(X_t, X_{t+1})dX_{t+1}.$$

These two forms can be grouped in the general expression

$$P(X_t|X_t) = q(X_{t+1}|X_t)\alpha(X_t, X_{t+1}) + I(X_{t+1} = X_t)[1 - \int q(X_{t+1}|X_t)\alpha(X_{t+1}|X_t)dX_{t+1}],$$

where $I(.)$ denotes the indicator function (taking the value 1 when its argument is true, and 0 otherwise).

In practice, simulation of a draw from π using the Metropolis-Hastings algorithm can be set up as follows:

Initialize X_0; set $t = 0$.

Repeat {

Sample a point Y from $q(.|X_t)$

Sample a Uniform(0,1) random varaibale U

If $U \leq \alpha(X_t, Y)$ set $X_{t+1} = Y$

otherwise set $X_{t+1} = X_t$

Increment t

}.

Random walk chains and symmetric $q(.|.)$

To implement the M-H algorithm, a suitable candidate-generating density should be specified. Typically, this density is selected from a family of distributions that require the specification of such tuning parameters as the location and scale. One family of candidate-generating densities, that appears in the work of Metropolis et al. (1953), is given by $q(X_t, X_{t+1}) = q_1(X_{t+1} - X_t)$, where $q_1(.|.)$ is a multivariate density. The candidate X_{t+1} is thus drawn according to the process $X_{t+1} = X_t + z$, where z is called the increment random variable and follows the distribution q_1. Because the candidate is equal to the current value plus noise, this case is called a *random walk* chain. Possible choices for q_1 include the multivariate normal density and the multivariate-t. Note that when q_1 is symmetric, the usual circumstance, $q_1(z) = q_1(-z)$; the probability of move then reduces to

$$\alpha(X, Y) = min\left(\frac{\pi(Y)}{\pi(X)}, 1\right).$$

There are some other family of candidate-generating densities. We are not discussing them here.

There is one critical issue of choosing the spread or scale of the candidate-generating density. This choice of scale certainly affects the efficiency of the algorithm and affects the behavior of the chain in at least two dimensions: one is the "acceptance rate" (the percentage of times a move to a new point is made), and the other is the region of the sample space that is covered by the chain. Consider the situation in which the chain has converged and the density is being sampled around the mode. Then, if the spread is extremely large, some of the generated candidates will be far from the current value, and will therefore have a low probability of being accepted. If the spread chosen is too small, the chain will take longer to traverse the support of the density, and stays in the low probability regions resulting in high acceptance rates. To get

a reasonable acceptance rate of around 50% we have to compromise between the two situations above.

1.5.4 Hybrid (HY) Algorithm

Sometimes simple random walk MCMC algorithms for estimating parameters of the high-dimensional target posterior density can not yield satisfactory estimates. To improve the algorithm the following two steps can be followed for a successful MCMC run: (a) incorporate derivative evaluations of the target log-posterior density, and (b) suppress the random walk behavior of the Markov chain. Incorporating derivative evaluations implies utilization of more information about the target distribution. By suppressing the random walk behavior we actually direct the chain to follow a definite path towards the target distribution for faster convergence and efficiency (see Neal [16],Gustafson et.al. [9]).

In the usual M-H algorithm we update estimates one at a time and have to set the jump size for each of the components of the parameter vector. In the guided random walk hybrid algorithm we can update components of k-dimensional parameter vector all at once and can set one jump size for the k-dimension. Here is the brief description of the general algorithm.

Let $X \sim \Pi$ be the target density having an unnormalized density function $\pi(x)$ on a subset of $\Re^k$. The algorithm works by extending the state from X to (X, Y), and the unnormalized target density from $\pi(x)$ to

$$\begin{aligned} \pi(x,y) &= \pi(x)\pi(y) \\ &= \pi(x)\exp\left(-\frac{1}{2}\sum_{i=1}^{k} y_i^2\right) \end{aligned} \tag{1.12}$$

where Y has a $N(0, I_k)$ distribution independent of X. Thus we can sample from $\pi(x)$ by sampling (X, Y) from (1.12) and simply discarding the Y value.

The following three steps should be followed to construct a Markov chain for (X, Y) having (1.12) as its stationary distribution. Also it is necessary to specify a step size $\epsilon > 0$, a function $g : \Re^k \to \Re^k$, and a constant $\delta \in [0, 1)$.

1. Determine a candidate state (x^*, y^*) as

$$\begin{aligned} x^* &\leftarrow x + \epsilon[y + (\epsilon/2)g(x)], \\ y^* &\leftarrow -y - (\epsilon/2)[g(x) + g(x^*)], \end{aligned}$$

and randomly assign

$$(x, y) \leftarrow \begin{cases} (x^*, y^*) & \text{with probability } p, \\ (x, y) & \text{with probability } 1 - p, \end{cases}$$

where

$$p = min\left\{\frac{\pi(x^*, y^*)}{\pi(x, y)}, 1\right\}.$$

2. Unconditionally negate y, i.e.

$$y \leftarrow -y.$$

3. Perform an autoregressive update to y, i.e.

$$y \leftarrow N(\delta y, (1 - \delta^2)^{1/2} I_k).$$

Thus we can improve our Monte Carlo sample estimates by using gradient evaluations $g(x) = \Delta \log \pi(x)$ of the parameters from the posterior distribution and should choose δ close to one to suppress random walk behavior.

1.6 Variable Selection

The goal of much research is to select those variables that results in the "best" predictive model when we have several potential independent variables to be included in the model. In many situations we have to consider interaction effects (say, pairwise) along with the main effects. Epidemiologists often suggest including all clinically and intuitively relevant variables in the model, regardless of their "statistical significance". But if the number of variables is large, it will be very difficult to get the actual effect of some biologically important variables. Thus, the approach should be seeking most parsimonious model that explains the data best. In order to achieve this goal we must have: (1) a basic plan for selecting variables, and (2) a set of methods for assessing the adequacy of the model.

There are several methods that one can follow to select variables for a logistic regression model. One method is "Univariate method" in which variables are assessed one by one via likelihood ratio test whether to include in the model. This method is time-consuming and tedious. Another approach is to use a "Stepwise method" in which variables are selected either

for inclusion or exclusion from the model in a sequential fashion based solely on statistical criteria. There are two main versions of the stepwise procedure: (a) forward selection with a test for backward elimination, and (b) backward elimination followed by a test for forward selection. In the next section we will describe briefly the first version and use it in our future variable-selection method. Hosmer and Lemeshow [12] describe the stepwise method and use likelihood ratio test to select a variable. We will use a different criterion, score test to select variables.

1.6.1 Stepwise Logistic Regression

In this method or algorithm a variable is selected on the basis of its "importance" in building up the model. The importance is defined in terms of a measure of the statistical significance of the coefficient for the variable and overall fit of the model. The statistic used is the score chi-square test. The specific procedure that we will use consists of the following few steps:

Step 0: Suppose we have available a total of p possible independent variables, all of which are candidates to be included in the model and are judged to have plausible "biological" importance in studying the response. Step (0) begins with a fit of the "intercept only model" and evaluation of Score chi-squares with corresponding p-values of all of the p factors. The first most important variable to be entered in the model is that which has the highest Score chi-square value and, of course, a small p-value.

A crucial aspect of implementing stepwise logistic regression is the choice of an reasonable "alpha(α)" level to judge the importance of variables. Let p_{E} be the choice where "E" stands for entry. The choice for p_{E} determines how many variables eventually are included in the model. Many researchers have studied the choice of p_{E} and their research has shown that $p_{\mathrm{E}} = 0.05$ is too stringent, often excluding important variables. Choosing a value for p_{E} in the range of 0.15 and 0.20 is highly recommended. If the goal of the analysis is broader and we want to include more variables that provides better prediction, we can choose $p_{\mathrm{E}} = 0.25$ or even larger. Whatever the choice of p_{E}, a variable is judged important enough to enter into the model if the p-value for Score chi-square is less than p_{E}.

Step 1: Step (1) begins with a fit of the logistic regression model containing the variable selected. The overall fit of the model is assessed via the likelihood ratio test and significance of

the coefficient of the maximum likelihood estimate is assessed.

To ascertain whether an entered variable should be deleted from the model the program selects that variable which, when removed, yields a high p-value. To decide whether the variable should be removed, the program compares the estimated p-value to second pre-chosen "alpha" level, p_{R}, which indicates some minimal level of continued contribution to the model where "R" stands for removal. Whatever value we choose for p_{R}, it must exceed the value of p_{E} to guard against the possibility of having the program enter and remove the same variable at successive steps. If we do not wish to exclude many variables once they have entered then we might use $p_{\mathrm{R}} = 0.9$. A more stringent value would be used if a continued "significant" contribution were required. For example, if we used $p_{\mathrm{E}} = 0.20$, then we might choose $p_{\mathrm{R}} = 0.25$. Thus if the p-value of the just entered variable exceeds p_{R} then the variable is removed from the model, otherwise it will stay in the model. The program again calculates score chi-square values for all remaining variables not in the model in this step.

Step 2: The procedure for step (2) is identical to that of step (1). The program selects the next variable to be entered into the model as the one having the highest score chi-square. Then it fits the model with the entered variables, assesses the overall fit of the model via the likelihood ratio test and assesses the significance of the maximum likelihood ratio estimates. Then it performs a check for the backward elimination. The process continues in this manner until the last variable selected according to p_{E} value. In the end, the process produces a summary table of maximum likelihood estimates of the variables selected for the final model.

The stepwise procedure that we will use in our model selection for different data sets will consist of the following two steps:

(a) In the first step we will select main effects only for our model using the stepwise procedure. We want to make sure that we are selecting significant effects to be used in the second step.

(b) In the second step we will ask the stepwise procedure to select different effects from among main effects (selected in the first step) and pairwise interactions of these main effects.

1.7 Cross-Validation

To compare the predictive power of Bayesian and classical models on some particular data sets, cross-validation is an effective method. The basic idea of cross-validation is to divide the whole data set randomly into a training sample and a validation or test sample. The results from this scheme may be sensitive to which particular subset of the data into training and test cases is utilized. For this reason, we use K-fold cross-validation in which we split the cases randomly into K roughly equal-sized segments. For example, when $K = 5$, the scenario looks like the following table:

1	2	3	4	5
Test	Train	Train	Train	Train

For the kth part (first in the table), we fit the model to the other $K - 1$ parts of the data, and calculate the predictive performance of the fitted model when predicting the kth part of the data. We do this for $k = 1, 2, \ldots, K$ and combine the K estimates of prediction. Each of the models to be compared is fit to the training sample first and then used to predict the test sample responses given the test sample covariates. Models which yield better predictions in this scheme are then preferred.

1.8 Outline and Scope of the Thesis

We begin Chapter 2 by discussing the pairwise interactions and how do these modify the effects. Then we introduce a different functional relationship between the response variable and the risk factors. We try to explain how to interpret the effects of particular risk factors under this new model. The Bayesian approach to parameter estimation and implementation of a specific MCMC method for the new model are discussed. In Chapter 3 we use some datasets from real epidemiological studies to fit the new model, interpret the parameter estimates, and compare the predictive power of the new model with that of the step-wise and ordinary logistic regression models. Chapter 4 discusses some limitations that we face in implementing the new model, and how we might overcome these restrictions in the future.

Chapter 2

Model Specification and Estimation

2.1 Interaction and Effect Modification

Generally, interaction is said to present between two risk factors when the effect of one risk factor upon disease is different at (at least some) different levels of the second risk factor. Consider a model containing a dichotomous risk factor (e.g. sex) and a continuous covariate (e.g. age). When interaction is present, the association between the risk factor and the outcome variable differs, or depends in some way on the level of the covariate. That is, the covariate modifies the effect of the risk factor. Epidemiologists often use the term *effect modifier* to describe a variable that interacts with a risk factor (see, for instance, Woodward [19]).

If the association between the covariate and the outcome variable is the same within each level of the risk factor, then there is no interaction between the covariate and the risk factor. Graphically, the absence of interaction yields a model with two parallel lines, one for each level of the risk factor (sex). In general, the absence of interaction is characterized by a model that contains no second or higher order terms involving two or more variables. In any epidemiologic study, we may have several risk factors that we decide to measure, and part of the study aims to decide which risk factors interact with others in regard to the disease outcome of interest. So then we can include in our model appropriate higher order terms to represent the effect of interaction.

An important step in the process of modeling a set of data is determining whether there is evidence of interaction in the data. If the number of risk factors in the data is relatively

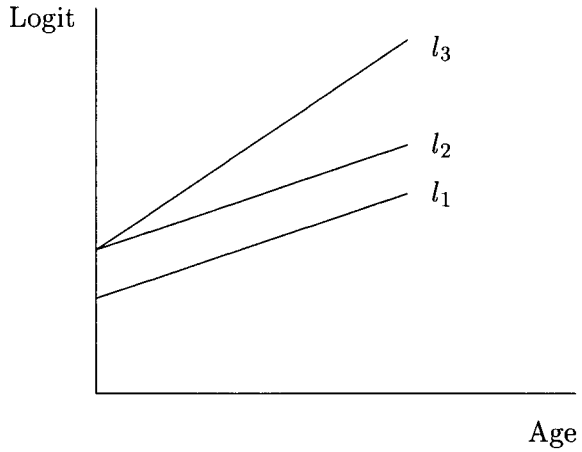

Figure 2.1: Plot of the logits under three different models showing the presence and absence of interaction

large, it's very difficult to identify which variables interact with each other.

Figure 2.1 presents the graphs of three different logits. The graphs of these logits will be used to explain what is meant by interaction. Consider an example where the outcome variable is the presence or absence of CHD, the risk factor is sex and the covariate is age. Suppose that the line l_1 corresponds to the logit for females as a function of age. Line l_2 represents the logit for males. These two lines are parallel to each other, indicating that the relationship between age and CHD is the same for males and females. In this situation there is no interaction and the log odds for sex (male versus female), controlling for age, is given by the differences between the lines l_2 and l_1. This difference is equal to the vertical distance between two lines, which is the same for all ages.

Suppose instead that the logit for males is given by the line l_3. This line is steeper than the line l_1, for females, indicating that the relationship between age and CHD among males is different from that among females. When this occurs we say there is an interaction between age and sex. The estimate of log-odds ratio for sex (male versus female) controlling for age is still given by the vertical distance between the lines $l_3 - l_1$, but this difference now depends on the age at which the comparison is made. Thus, we can not estimate the odds ratio for sex without specifying the age. That is, age is an effect modifier.

To consider the magnitude of "effect modification", say that CHD depends on two risk

factors, smoking status and systolic blood pressure. For simplicity, we assume that both risk factors are binary and can be defined as:

$$X_1 = \begin{cases} 0 & \text{if no smoking,} \\ 1 & \text{if smoking,} \end{cases}$$

and

$$X_2 = \begin{cases} 0 & \text{if blood pressure is normal,} \\ 1 & \text{if blood pressure is high.} \end{cases}$$

Let β_1 and β_2 represent the coefficients of X_1 and X_2, respectively. Then we can express the relationship between the probability of presence of CHD and the risk factors in the logit scale as (assuming interaction effect of smoking and blood pressure is present):

$$logit\ Pr(Y = 1|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2, \tag{2.1}$$

where β_{12} is the magnitude of effect modification. Say that the effect of smoking is of interest. Comparing $X_1 = 0$ and $X_1 = 1$ we have

Level of X_2	logit for $X_1 = 0$	logit for $X_1 = 1$	logit difference
$X_2 = 0$	β_0	$\beta_0 + \beta_1$	β_1
$X_2 = 1$	$\beta_0 + \beta_2$	$\beta_0 + \beta_1 + \beta_2 + \beta_{12}$	$\beta_1 + \beta_{12}$

giving a sense in which β_{12} describes how the smoking effect is modified by blood pressure. Similarly, it can be shown that the blood pressure effect is modified by smoking, by the amount β_{12}, if we quantify the blood pressure effect by comparing $X_2 = 0$ and $X_2 = 1$ for two levels of smoking.

Suppose we have 10 risk factors to consider. If we want to include all of them in the model without considering their statistical significance, and also want to evaluate their pairwise interaction effects, we will have 10+45=55 terms to include in the model. The additive model will then look like:

$$logit\ Pr(Y = 1|X) = \beta_0 + \beta_1 X_1 + \ldots + \beta_{10} X_{10} + \beta_{1,2} X_1 X_2 + \ldots + \beta_{9,10} X_9 X_{10}. \tag{2.2}$$

In such a situation most of the parameter estimates will be insignificant and unreliable due to confounding or some other factors, and evaluation of interaction effects will be unreliable. This

is not a feasible way of evaluating effects. The most widely used procedure for selecting main effects and interaction effects is the stepwise logistic regression procedure discussed in Chapter 1, Section 1.4. Nevertheless, one might raise several issues such as:

1. Will a pairwise interaction model always be realistic ?
2. Is an additive structure as in (2.1) always appropriate ?
3. Could there be other kinds of interaction whereby the effect of a given predictor is different for a subject at generally low risk than for a subject at generally high risk ?

In an attempt to discuss these issues, suppose we have several risk factors under study and one of them is smoking status. Say we want to consider the effect of smoking on CHD as a function of the remaining predictors. Even though there may be numerous other risk factors, hypothetically consider the X-axis in Figure 2.2 to summarize their combined effect. If it were possible to place these predictors on this X-axis of the two-dimensional plane, from low levels to high levels, and smoking effect on the Y-axis, then we can contemplate the following figure. This picture is trying to grasp the idea of issue 3 above. From this figure we see that the effect

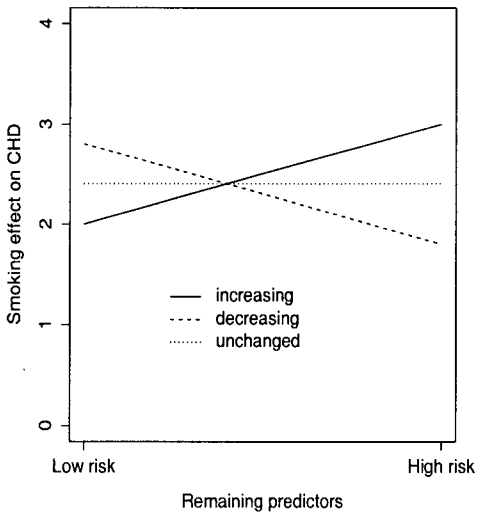

Figure 2.2: Hypothetical representation of relationship among risk factors.

of smoking on CHD, considering remaining predictors from their low risk level to high risk level, might be constant, increasing or decreasing. The ordinary interaction model (2.1) is not amenable to describing this situation. So we wish to consider alternate models. We will return to this issue in Section 2.2.

To get some sense of other issues, let us rewrite equation (2.1) as:

$$logit\ Pr(Y = 1|X) = \beta_0 + g(X_1 + \ldots + X_p), \tag{2.3}$$

where p components of X are the risk factors and g is the function of the risk factors. In equation (2.1) we assumed g as linear and assumed additive effect of the risk factors on the disease. It may not always be true. It might be possible that after reaching a certain level, the risk on the disease outcome will be almost constant and effects of additional risk factors will be low. This situation encourages us to think g as a non-linear function of the risk factors. The following figure illustrates plausible non-linearity of the logit as a function of the number of risk factors present:

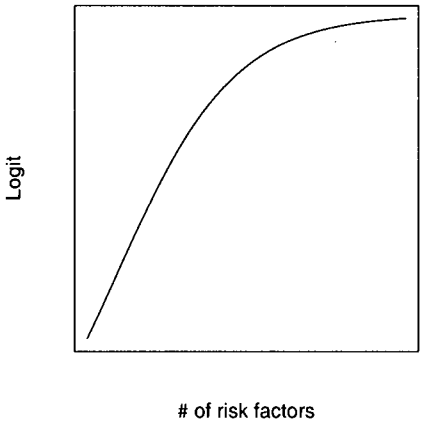

Figure 2.3: Logit as non-linear function of the number of risk factors

Say risk of the disease is the function of the number of risk factors where all of them are binary, representing e.g. absence and presence of the risk factors. Thus risk of the disease in the logit scale might be non-linear as plotted in Figure 2.3 above. From this figure, it can

be said that as the number of risk factor increases, the increased risk of the disease on the logit scale will be getting smaller and smaller. Qualitatively, we can conclude that the effect of an additional risk factor will be small as the number of risk factor increases, even if the risk factor may be biologically important. More precisely, let Z denote the number of risk factors. Then unit increase in Z will have an insignificant effect on the logit if Z is large. For making the above comments we assume here that all of the risk factors are equally important. No single risk factor has greater influence on the risk than any other. A new model will be introduced in Section 2.2 to generalize this behavior beyond the special case of equally important risk factors.

In conclusion, instead of trying to identify the possible pairwise interaction effects for fitting the linear model, we could try to find some tractable and interpretable functional form of g. What functional form g could take? Could g interpret the situations we discussed above? In the next section a specific functional form of g will be assumed and will be used to describe the relationship between the logit and the risk factors.

2.2 Specific non-additive functional form of the logit

Suppose we have $X_i,\ i = 1, 2, \ldots, p$ predictors where :

$$X_i = \begin{cases} 0 & \text{if the } i\text{-th risk factor is absent,} \\ 1 & \text{if the } i\text{-th risk factor is present.} \end{cases}$$

The specific non-linear functional form of g that we are thinking of can be expressed in terms of logit and is given by:

$$logit\ Pr(Y = 1|X) = \beta_0 + \left\{\beta_1^{\lambda} X_1 + \ldots + \beta_p^{\lambda} X_p\right\}^{1/\lambda}, \tag{2.4}$$

where λ is an additional unknown parameter and X_i's are binary as defined above. We call this model (2.4) as **NAD** (Non-ADditive) model throughout the entire thesis. The NAD model is not defined properly unless we restrict the signs of the coefficients β_i to be positive. This implies that $X_i = 1$ corresponds to higher risk than $X_i = 0$. The functional form may have some advantages over trying to pick out pairwise interaction terms.

How do we interpret λ? Consider a very simple case: $\beta_1 = \beta_2 = \ldots = \beta_p \equiv \beta > 0$. Then

equation (2.4) reduces to:

$$\begin{aligned} \textit{logit } Pr(Y=1|X) &= \beta_0 + \left\{\beta^\lambda (X_1 + \ldots + X_p)\right\}^{1/\lambda}, \\ &= \beta_0 + \beta(\#\text{ of risk factors})^{1/\lambda}. \end{aligned} \tag{2.5}$$

In this special setting, we see from Equation (2.5) that logit depends on how many risk factors a person has and obviously on the value of λ. This special behavior of the logit where each of the risk factors is getting the same weight, is represented in Figure 2.3.

To interpret β by the odds ratio let us compare the logit of somebody having no risk factors with the logit of a person with just one risk factor (the i-th one). Using Equation (2.5) we have:

Risk factor	0	i-th one
logit	β_0	$\beta_0 + \beta_i$

Thus the odds ratio is e^{β_i}, just like the ordinary logistic regression model and bears the same interpretation.

Now let $p = 10$ and $\beta = 1$. From Equation (2.5) we see that the functional form of logit depends on the value of λ. For $\lambda = 1$ we get the simple linear logistic function. For other values of λ we can depict logit as a function of number of risk factors in this special example. This is shown in the first picture of Figure 2.4.

What can be said about the effect of an additional risk factor on the risk for someone who has already p risk factors? For $\lambda = 1$ it would be constant. For $\lambda > 1$ and $\lambda < 1$ the effect would be decreasing and increasing, respectively. This is depicted in the second picture of Figure 2.4.

This simple example motivates us to think that logit might be a non-linear function of the risk factors with λ different from one. But does a value of λ different from one tend to explain the data better? How interpretable are such models? We will discuss these issues using some real life examples after estimating λ and the coefficient β's.

However, in reality the usual picture is that β's are not all equal. Let $X = (X_i, X_{(i)})$, where $X_{(i)}$ represent factors other than X_i. Suppose the effect of X_i is of interest and could be quantified by:

$$\text{logit}(X_i = 1, X_{(i)}) - \text{logit}(X_i = 0, X_{(i)}). \tag{2.6}$$

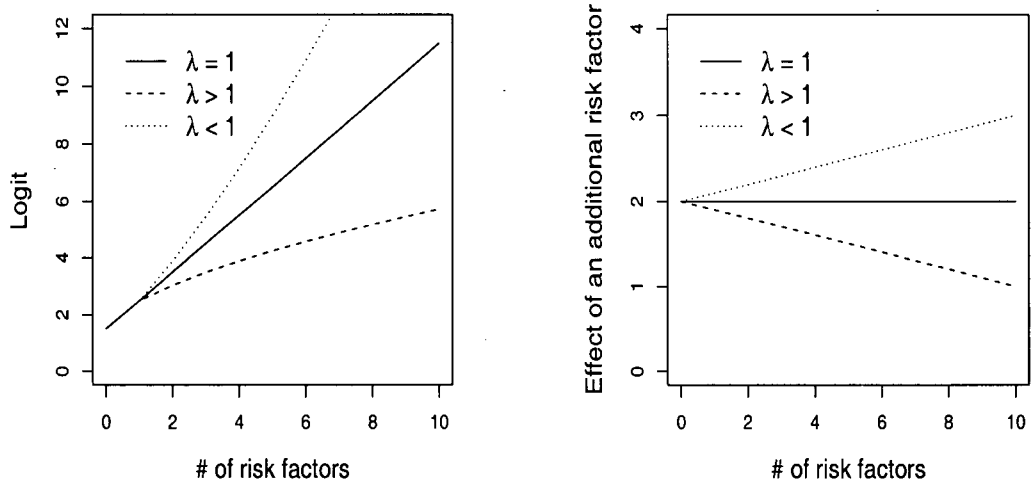

Figure 2.4: Comparison of different λ values.

Using the NAD model effect of X_i can be assessed from equation (2.6) as follows:

(i) If $\lambda > 1$ then (2.6) is decreasing in each $X_j,\ j \neq i$.

(ii) If $\lambda = 1$ then (2.6) is unaffected by $X_{(i)}$.

(iii) If $\lambda < 1$ then (2.6) is increasing in each $X_j,\ j \neq i$.

Which value of λ is best supported by data? How well does the NAD model describe the data? To find the answers we have to fit the model first and estimate the parameters.

2.3 Model Fitting and Estimation

We assumed that we have $y_1, y_2, \ldots, y_n$ independent binary outcomes of a disease of n independent observations in any study. For each of the y_i's we have corresponding $X_1, X_2, \ldots, X_p$ predictors. More formally, we can define our observation as the pair (y_i, x_i) for $i = 1, 2, \ldots, n$ where $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$.

2.3.1 Log Likelihood Function

Let us denote the NAD model by

$$g(X;\beta,\lambda) \equiv logit\ Pr(Y=1|X) = \beta_0 + \left\{\beta_1^\lambda X_1 + \ldots + \beta_p^\lambda X_p\right\}^{1/\lambda}. \tag{2.7}$$

Then we express the logistic regression model (2.7) as

$$Pr(Y=1|X=x) = \frac{e^{g(X;\beta,\lambda)}}{1+e^{g(X;\beta,\lambda)}},$$

or

$$\pi(x_i) = \frac{e^{\beta_0+\left(\sum_{j=1}^{p}\beta_j^\lambda x_{ij}\right)^{1/\lambda}}}{1+e^{\beta_0+\left(\sum_{j=1}^{p}\beta_j^\lambda x_{ij}\right)^{1/\lambda}}}. \tag{2.8}$$

Therefore, the likelihood function can be written in the following form:

$$L(\beta,\lambda) = \prod_{i=1}^{n}\pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i}.$$

Rearranging and following the procedure of Section 1.2.1, Chapter 1, it turns out that the likelihood function equals:

$$L(\beta,\lambda) = \left\{\prod_{i=1}^{n}\frac{1}{\left[1+\exp\left(\beta_0+\left(\sum_{j=1}^{p}\beta_j^\lambda x_{ij}\right)^{1/\lambda}\right)\right]}\right\} \times \exp\left[\sum_{i=1}^{n} y_i\left\{\beta_0+\left(\sum_{j=1}^{p}\beta_j^\lambda x_{ij}\right)^{1/\lambda}\right\}\right] \tag{2.9}$$

which yields the log likelihood function given by:

$$l(\beta,\lambda) = \sum_{i=1}^{n}\left[y_i\left\{\beta_0+\left(\sum_{j=1}^{p}\beta_j^\lambda x_{ij}\right)^{1/\lambda}\right\} - \log\left[1+\exp\left(\beta_0+\left(\sum_{j=1}^{p}\beta_j^\lambda x_{ij}\right)^{1/\lambda}\right)\right]\right] \tag{2.10}$$

When $\lambda = 1$, equation (2.10) reduces to the log likelihood function of the ordinary logistic regression model. In the ordinary logistic regression case we use Maximum Likelihood Principle to estimate the parameters β. To estimate the parameters of the NAD model we will follow the Bayesian approach.

2.3.2 Posterior density

The parameter λ is positive. We will reparameterize λ by defining $\phi = \log\lambda$ for numerical simplicity. So $\lambda = \exp(\phi)$ and $L(\beta, \lambda)$ becomes $L(\beta, \phi)$.

We assume a normal prior for ϕ, i.e. $\pi(\phi) \sim N(0, c^2)$, or

$$\pi(\phi) = \frac{1}{c\sqrt{2\pi}} \exp\left(-\frac{1}{2c^2}\phi^2\right).$$

To avoid some numerical complexities associated with large λ values we assume $c = \log(2)$. By choosing a symmetric prior for ϕ to be centered at 0, λ and $\frac{1}{\lambda}$ are equally likely *a priori*. Centering the prior for ϕ at 0 corresponds to centering the prior for λ at 1, thereby favouring the simple logistic regression model.

For parameters β we assume a noninformative diffuse prior distribution, that is $\pi(\beta) \propto 1$. In practice, we might use a uniform prior distribution if we really have no prior knowledge about the parameters. The joint posterior distribution of all the parameters given the data is then defined as:

$$P(\beta, \phi|y) \propto L(\beta, \phi)\pi(\beta)\pi(\phi).$$

Therefore we can get the log posterior density as follows:

$$\begin{aligned} \log P(\beta, \phi|y) &= K + \sum_{i=1}^{n} y_i \left\{ \beta_0 + \left(\sum_{j=1}^{p} \beta_j^{e^\phi} x_{ij} \right)^{e^{-\phi}} \right\} \\ &- \sum_{i=1}^{n} \log \left[1 + \exp\left(\beta_0 + \left(\sum_{j=1}^{p} \beta_j^{e^\phi} x_{ij} \right)^{e^{-\phi}} \right) \right] - \frac{1}{2c^2}\phi^2, \end{aligned} \tag{2.11}$$

where K is an unknown constant.

In the next section we will discuss how to estimate the parameters from this density by Markov Chain Monte Carlo methods.

2.3.3 Parameter estimation

In the Bayesian perspective, we get estimates of the parameters from their posterior means as we discussed in Chapter 1, Section 1.2.2. If we want to estimate β's, e.g. β_1, we have to evaluate

Equation (1.9) using the posterior density (2.11). One way to evaluate (1.9) is to compute conditional distributions of the parameters from the density (2.11) and then draw samples from their respective conditional distributions. This is known as the Gibbs sampler. But the conditional distributions of the parameters from density (2.11) do not have nice and simple mathematical closed form. Alternatively, we can draw samples from the posterior density (2.11) by the Random Walk Metropolis-Hastings (MH) algorithm discussed in Chapter 1, Section 1.3.3. Thus we can draw m samples of ϕ, for example, and get $\lambda = \exp(\phi)$. We can then estimate $\hat{\lambda} = \frac{\lambda^{(1)}+\ldots+\lambda^{(m)}}{m}$ to approximate $\int \lambda \log P(\beta, \lambda|y) d\beta d\lambda$.

To implement MH algorithm and simulate samples from Equation (2.11) using some data we need to specify:

(i) A candidate generating density: We will use Normal density centered at the current parameter value with specific standard deviations or jump sizes.

(ii) Initial values of β and ϕ. We will use estimates from the ordinary logistic regression fit of the data as initial values for β, and $\phi^0 = 0$.

(iii) Jump sizes for each β and ϕ.

An important step in simulation is the setting up of jump sizes for each parameter, for proper mixing and a reasonable acceptance rate. We tried to draw samples from (2.11) using a data set (will be discussed in the next section) through RW MH algorithm. But due to high correlation between ϕ and β we were having trouble with the mixing of the sampler output, with a very low acceptance rate after several changes of the jump sizes. We then tried to rescale the β's by $\tilde{\beta} = \frac{\beta}{\exp(\phi)}$ while fixing the value of ϕ. Rescaling was implemented through multiplying jump sizes by $\exp(\phi)$. The mixing was good and acceptance rate was satisfying, but we do not have estimates of ϕ. Besides, there's no established or specific rule to specify jump sizes. They are fixed by trial and error basis. An alternative way to get simultaneous estimates of the parameters is to use Hybrid (HY) algorithm by using full information of the log posterior density which was discussed in Chapter 1, Section 1.3.4. Moreover, we can get rid of specifying several jump sizes by specifying only one jump size for all of the parameters.

To implement the HY algorithm we need to evaluate the derivatives of (2.11) with respect

to parameters β and ϕ. Rewrite Equation (2.11) as follows:

$$\log P(\beta,\phi|y) = K + \sum_{i=1}^{n}\left[y_i g_i(\beta,\phi) - \log(1+e^{g_i(\beta,\phi)})\right] - \frac{1}{2c^2}\phi^2, \tag{2.12}$$

where

$$g_i(\beta,\phi) = \beta_0 + \left(\sum_{j=1}^{p}\beta_j^{e^\phi} X_{ij}\right)^{e^{-\phi}}. \tag{2.13}$$

Differentiating (2.12) with respect to $\beta_j, j = 0,1,2,\ldots,p$ and ϕ we get the following $(p+2)$ equations:

$$\begin{aligned}
\frac{\partial \log P(\beta,\phi|y)}{\partial \beta_j} &= \sum_{i=1}^{n}\left[y_i\left(\frac{\partial}{\partial\beta_j}g_i(\beta,\phi)\right) - \frac{e^{g_i(\beta,\phi)}}{1+e^{g_i(\beta,\phi)}} \times \frac{\partial}{\partial\beta_j}g_i(\beta,\phi)\right] \\
&= \sum_{i=1}^{n}\left[y_i - \left(1 - \frac{1}{1+e^{g_i(\beta,\phi)}}\right)\right]\frac{\partial}{\partial\beta_j}g_i(\beta,\phi), \qquad (2.14) \\
\frac{\partial \log P(\beta,\phi|y)}{\partial \phi} &= \sum_{i=1}^{n}\left[y_i - \left(1 - \frac{1}{1+e^{g_i(\beta,\phi)}}\right)\right]\frac{\partial}{\partial\phi}g_i(\beta,\phi) - \frac{\phi}{c^2}. \qquad (2.15)
\end{aligned}$$

To get $(p+1)$ equations of (2.14) we need to evaluate $\frac{\partial}{\partial\beta_j}g_i(\beta,\phi)$. Differentiating (2.13) with respect to β_j for $j = 0,1,2,\ldots,p$ we get the following $(p+1)$ equations:

$$\begin{aligned}
\frac{\partial g_i(\beta,\phi)}{\partial\beta_0} &= 1, \\
\frac{\partial g_i(\beta,\phi)}{\partial\beta_1} &= \frac{\partial}{\partial\beta_1}\exp\left(e^{-\phi}\log\left(\sum_{j=1}^{p}\beta_j^{e^\phi}X_{ij}\right)\right) \\
&= \left(\sum_{j=1}^{p}\beta_j^{e^\phi}X_{ij}\right)^{e^{-\phi}}\frac{\partial}{\partial\beta_1}\left[e^{-\phi}\log\left(\sum_{j=1}^{p}\beta_j^{e^\phi}X_{ij}\right)\right] \\
&= \left(\sum_{j=1}^{p}\beta_j^{e^\phi}X_{ij}\right)^{e^{-\phi}} e^{-\phi}\times\frac{1}{\left(\sum_{j=1}^{p}\beta_j^{e^\phi}X_{ij}\right)}\frac{\partial}{\partial\beta_1}\left(\sum_{j=1}^{p}\beta_j^{e^\phi}X_{ij}\right) \\
&= \left(\sum_{j=1}^{p}\beta_j^{e^\phi}X_{ij}\right)^{e^{-\phi}-1} e^{-\phi}e^{\phi}\beta_1^{e^\phi-1}X_{i1} \\
&= \left(\sum_{j=1}^{p}\beta_j^{e^\phi}X_{ij}\right)^{e^{-\phi}-1} \beta_1^{e^\phi-1}X_{i1}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\frac{\partial g_i(\beta,\phi)}{\partial \beta_2} &= \left(\sum_{j=1}^{p} \beta_j^{e^\phi} X_{ij}\right)^{e^{-\phi}-1} \beta_2^{e^\phi-1} X_{i2}, \\
&\vdots \\
\frac{\partial g_i(\beta,\phi)}{\partial \beta_p} &= \left(\sum_{j=1}^{p} \beta_j^{e^\phi} X_{ij}\right)^{e^{-\phi}-1} \beta_p^{e^\phi-1} X_{ip}.
\end{aligned}$$

Equation (2.14) can now be evaluated using the above equations. Further we need to evaluate $\frac{\partial}{\partial \phi} g_i(\beta,\phi)$ which is given by,

$$\begin{aligned}
\frac{\partial g_i(\beta,\phi)}{\partial \phi} &= \frac{\partial}{\partial \phi} \exp\left(e^{-\phi} \log\left(\sum_{j=1}^{p} \beta_j^{e^\phi} X_{ij}\right)\right) \\
&= \left(\sum_{j=1}^{p} \beta_j^{e^\phi} X_{ij}\right)^{e^{-\phi}} \frac{\partial}{\partial \phi}\left[e^{-\phi} \log\left(\sum_{j=1}^{p} \beta_j^{e^\phi} X_{ij}\right)\right] \\
&= \left(\sum_{j=1}^{p} \beta_j^{e^\phi} X_{ij}\right)^{e^{-\phi}} \left[e^{-\phi} \frac{\partial}{\partial \phi} \log\left(\sum_{j=1}^{p} \beta_j^{e^\phi} X_{ij}\right) + \log\left(\sum_{j=1}^{p} \beta_j^{e^\phi} X_{ij}\right) \frac{\partial}{\partial \phi} e^{-\phi}\right] \\
&= \left(\sum_{j=1}^{p} \beta_j^{e^\phi} X_{ij}\right)^{e^{-\phi}} \left[\frac{e^{-\phi}}{\left(\sum_{j=1}^{p} \beta_j^{e^\phi} X_{ij}\right)} \frac{\partial}{\partial \phi}\left(\sum_{j=1}^{p} \beta_j^{e^\phi} X_{ij}\right) + \right. \\
&\qquad \left. \log\left(\sum_{j=1}^{p} \beta_j^{e^\phi} X_{ij}\right)(-e^{-\phi})\right] \\
&= \left(\sum_{j=1}^{p} \beta_j^{e^\phi} X_{ij}\right)^{e^{-\phi}} \left[\frac{e^{-\phi}}{\left(\sum_{j=1}^{p} \beta_j^{e^\phi} X_{ij}\right)} \left(\sum_{j=1}^{p} \frac{\partial}{\partial \phi}(\beta_j^{e^\phi}) X_{ij}\right) - \right. \\
&\qquad \left. e^{-\phi} \log\left(\sum_{j=1}^{p} \beta_j^{e^\phi} X_{ij}\right)\right].
\end{aligned}$$

Now, since

$$\begin{aligned}\frac{\partial}{\partial\phi}(\beta_j^{e^\phi}) &= \frac{\partial}{\partial\phi}\exp\left(e^\phi\log(\beta_j)\right)\\ &= \beta_j^{e^\phi}\frac{\partial}{\partial\phi}\left(e^\phi\log(\beta_j)\right)\\ &= \beta_j^{e^\phi}\left(e^\phi\log(\beta_j)\right),\end{aligned}$$

therefore,

$$\begin{aligned}\frac{\partial g_i(\beta,\phi)}{\partial\phi} &= \left(\sum_{j=1}^p\beta_j^{e^\phi}X_{ij}\right)^{e^{-\phi}}\left[\frac{e^{-\phi}}{\left(\sum_{j=1}^p\beta_j^{e^\phi}X_{ij}\right)}e^\phi\sum_{j=1}^p\left(\beta_j^{e^\phi}\log(\beta_j)X_{ij}\right)-\right.\\ &\qquad\left.e^{-\phi}\log\left(\sum_{j=1}^p\beta_j^{e^\phi}X_{ij}\right)\right]\\ &= \left(\sum_{j=1}^p\beta_j^{e^\phi}X_{ij}\right)^{e^{-\phi}}\left[\frac{\sum_{j=1}^p\left(\beta_j^{e^\phi}\log(\beta_j)X_{ij}\right)}{\left(\sum_{j=1}^p\beta_j^{e^\phi}X_{ij}\right)}-e^{-\phi}\log\left(\sum_{j=1}^p\beta_j^{e^\phi}X_{ij}\right)\right].\end{aligned}$$

Even though after using derivative information of the log posterior density in the HA we were not getting satisfactory sample estimates, that is, mixing of the sampler output was not good. We then redefine $\phi = k\log\lambda$, where k is a constant, then $\lambda = e^{\phi/k}$. The prior distribution of ϕ now depends on k. Since $\phi/k \sim N(0,c^2)$, $\phi \sim N(0,c^2k^2)$. We assume $k=4$ and $c = \frac{1}{2}\log(2)$. Still we have not got reasonable acceptance rate and proper mixing of the estimates due to correlation between ϕ and β's. We then reparameterized β's as $\alpha_j = \frac{\beta_j}{\lambda}$ for $j = 1,2,\ldots,p$ so that $\beta_j = \lambda\alpha_j$. In this reparameterization we have to redefine the NAD model as follows:

$$\begin{aligned}logit\ Pr(Y=1|X) &= \beta_0+\left[(\lambda\alpha_1)^\lambda X_1+\ldots+(\lambda\alpha_p)^\lambda X_p\right]^{1/\lambda}\\ &= \beta_0+\left[\lambda^\lambda(\alpha_1^\lambda X_1+\ldots+\alpha_p^\lambda X_p)\right]^{1/\lambda}\\ &= \alpha_0+\lambda\left[\alpha_1^\lambda X_1+\ldots+\alpha_p^\lambda X_p\right]^{1/\lambda},\end{aligned}\tag{2.16}$$

where $\alpha_0 = \beta_0$. Let

$$g_i(\alpha,\phi) = \alpha_0+e^{(\phi/k)}\left[\alpha_1^{e^{(\phi/k)}}X_{i1}+\ldots+\alpha_p^{e^{(\phi/k)}}X_{ip}\right]^{e^{-(\phi/k)}}.\tag{2.17}$$

Then the log posterior density is given by:

$$\log P(\alpha, \phi|y) = K + \sum_{i=1}^{n} \left[y_i g_i(\alpha, \phi) - \log(1 + e^{g_i(\alpha,\phi)}) \right] - \frac{1}{2c^2k^2}\phi^2. \tag{2.18}$$

Differentiating (2.17) with respect to α_j for $j = 0, 1, 2, \ldots, p$ we get the following $(p+1)$ equations:

$$\begin{aligned}
\frac{\partial g_i(\alpha,\phi)}{\partial \alpha_0} &= 1, \\
\frac{\partial g_i(\alpha,\phi)}{\partial \alpha_1} &= e^{(\phi/k)} \left(\sum_{j=1}^{p} \alpha_j^{e^{(\phi/k)}} X_{ij} \right)^{e^{-(\phi/k)}-1} \alpha_1^{e^{(\phi/k)}-1} X_{i1}, \\
\frac{\partial g_i(\alpha,\phi)}{\partial \alpha_2} &= e^{(\phi/k)} \left(\sum_{j=1}^{p} \alpha_j^{e^{(\phi/k)}} X_{ij} \right)^{e^{-(\phi/k)}-1} \alpha_2^{e^{(\phi/k)}-1} X_{i2}, \\
&\vdots \\
\frac{\partial g_i(\alpha,\phi)}{\partial \alpha_p} &= e^{(\phi/k)} \left(\sum_{j=1}^{p} \alpha_j^{e^{(\phi/k)}} X_{ij} \right)^{e^{-(\phi/k)}-1} \alpha_p^{e^{(\phi/k)}-1} X_{ip}.
\end{aligned}$$

Equation (2.14) can now be evaluated using these equations in α parameterization. Further $\frac{\partial}{\partial \phi} g_i(\alpha, \phi)$ can be obtained as follows:

$$\begin{aligned}
\frac{\partial g_i(\alpha,\phi)}{\partial \phi} &= \frac{e^{(\phi/k)}}{k} \left(\sum_{j=1}^{p} \alpha_j^{e^{(\phi/k)}} X_{ij} \right)^{e^{-(\phi/k)}} \\
&\times \left[1 + \left(\frac{\sum_{j=1}^{p} \alpha_j^{e^{(\phi/k)}} \log(\alpha_j) X_{ij}}{\left(\sum_{j=1}^{p} \alpha_j^{e^{(\phi/k)}} X_{ij} \right)} - e^{-(\phi/k)} \log \left(\sum_{j=1}^{p} \alpha_j^{e^{(\phi/k)}} X_{ij} \right) \right) \right].
\end{aligned}$$

Equation (2.15) can be evaluated using this equation where the last term would be $\frac{\phi}{c^2k^2}$.

Note that these above expressions of gradients do not cover the situation if all of the risk factors of a person have value zero. In this situations model (2.16) reduces to *logit* $Pr(Y =$

$1|X) = \alpha_0$ and Equation (2.17) reduces to $g_i(\alpha, \phi) = \alpha_0$. We then have the derivatives as:

$$\begin{aligned} \frac{\partial g_i(\alpha,\phi)}{\partial \alpha_0} &= 1 \\ \frac{\partial g_i(\alpha,\phi)}{\partial \alpha_1} &= \frac{\partial g_i(\alpha,\phi)}{\partial \alpha_2} = \ldots = \frac{\partial g_i(\alpha,\phi)}{\partial \alpha_p} = 0, \end{aligned}$$

and

$$\frac{\partial g_i(\alpha,\phi)}{\partial \phi} = 0.$$

So we will not get any information about the coefficients and λ except the intercept. We have to keep it in mind during the MCMC run.

The next chapter will be devoted to estimate the parameters of the model (2.16) by using some real data sets, where we compare the performance of the NAD model with that of the step-wise and the ordinary logistic regression model.

Chapter 3

Examples

3.1 Example 1 : South African Heart Study (SAHS)

This dataset contains a retrospective sample of 460 males in a heart-disease high-risk region of the Western Cape, South Africa. The response variable was Coronary Heart Disease (CHD) represented as presence and absence of the disease. There are roughly two controls per case of CHD. Many of the CHD positive men have undergone blood pressure reduction treatment and other programs to reduce their risk factors after their CHD event. In some cases the measurements were made after these treatments. Of the 460 males there were 160 people who had the CHD event. There were nine risk factors measured in this study; one of them is binary and the remaining eight are continuous. Complete descriptions for some of the predictors were not available. These data are taken from a larger dataset, described in Rousseauw et. al. [17]. In Table 3.1 the names, description and type of the risk factors are given.

We convert the eight continuous risk factors into binary variables by thresholding in comparison to their means.

For evaluating the Hybrid (HY) Algorithm and to get a reasonable acceptance rate while ensuring proper mixing of the sampler output, we assume $k = 4, c = \frac{1}{2}\log(2), \epsilon = 0.07, \delta = 0.90$. As we have mentioned before the initial values for β's were the estimates from the ordinary logistic regression fit and $\phi^0 = 0$. After iterating the algorithm for 20,000 times the acceptance rate was 82% and the Markov Chain is stabilized as can be seen from Figure 3.1. Figure 3.1 and 3.2 show the sample path of the MCMC output and the posterior distribution by histogram,

Table 3.1: Description of the risk factors for SAHS

Variable Name	Description	Type
SBP	Systolic Blood Pressure	Continuous
Tobacco	Cumulative tobacco consumption (kg)	Continuous
LDL	Low density lipoprotein cholesterol	Continuous
Adiposity*	Measured fatness	Continuous
Typea	Type-A behavior	Continuous
Obesity*	Measured fatness	Continuous
Alcohol	Current alcohol consumption	Continuous
Age	Age at onset	Continuous
FamHist	Family history of heart disease (Present, Absent)	Binary

*Incomplete information.

respectively.

By looking at the plots in Figure 3.1 we see the almost instantaneous convergence of the sampler to the target posterior distribution. For this reason we do not throw away any 'burn-in' iterations to compute the posterior means and standard deviations. As a summary of the posterior distribution given the data, the posterior means and the corresponding posterior standard deviations, computed using the 20,000 iterations, are given in Table 3.2. The 90% equal-tailed credible intervals are also computed from the sample quantiles of the MCMC output, presented in Table 3.2. The predictors in Table 3.2 are listed in the order they were given in Table 3.1.

The restriction imposed on the coefficients to be positive has some effect on some of the posterior distributions of the parameters as can be seen from the histograms in Figure 3.2 .

We now fit the ordinary logistic regression model to the data to compare with the NAD model. The estimates, standard errors and the 95% confidence intervals are given in Table 3.3.

Comparing the estimates from Table 3.3 with those MCMC output from the NAD model we see that the NAD model tends to provide larger estimated β_i's than does the ordinary logistic regression model and the posterior standard deviations are much bigger than the standard errors estimated by the ordinary logistic regression model. It is quite reasonable to have bigger posterior standard deviations since while sampling from the posterior density of the NAD model we incorporate more uncertainty through the additional parameter λ and it is evident that data

Table 3.2: Summary results of the posterior distribution

	Posterior		90% CI	
Variable	Means	Std Dev.	5%	95%
Intercept	-3.91	0.83	-5.53	-2.76
X_1	1.11	0.69	0.18	2.47
X_2	1.13	0.66	0.24	2.37
X_3	0.92	0.60	0.11	2.07
X_4	1.18	0.74	0.18	2.53
X_5	1.28	0.69	0.33	2.58
X_6	0.94	0.68	0.09	2.28
X_7	0.51	0.44	0.02	1.36
X_8	1.75	0.71	0.81	3.06
X_9	1.87	0.72	0.88	3.27
λ	1.52	0.34	1.06	2.19

Table 3.3: Summary results from the ordinary logistic regression fit

			95% CI	
Coefficients	Estimate	Std Error	Lower	Upper
Intercept	-2.70	0.37	-3.44	-1.97
X_1	0.38	0.23	-0.07	0.83
X_2	0.48	0.23	0.02	0.94
X_3	0.38	0.23	0.07	0.84
X_4	0.34	0.30	-0.26	0.93
X_5	0.52	0.22	0.08	0.95
X_6	0.21	0.27	-0.32	0.74
X_7	0.06	0.23	-0.40	0.51
X_8	0.87	0.26	0.36	1.39
X_9	0.93	0.22	0.50	1.36

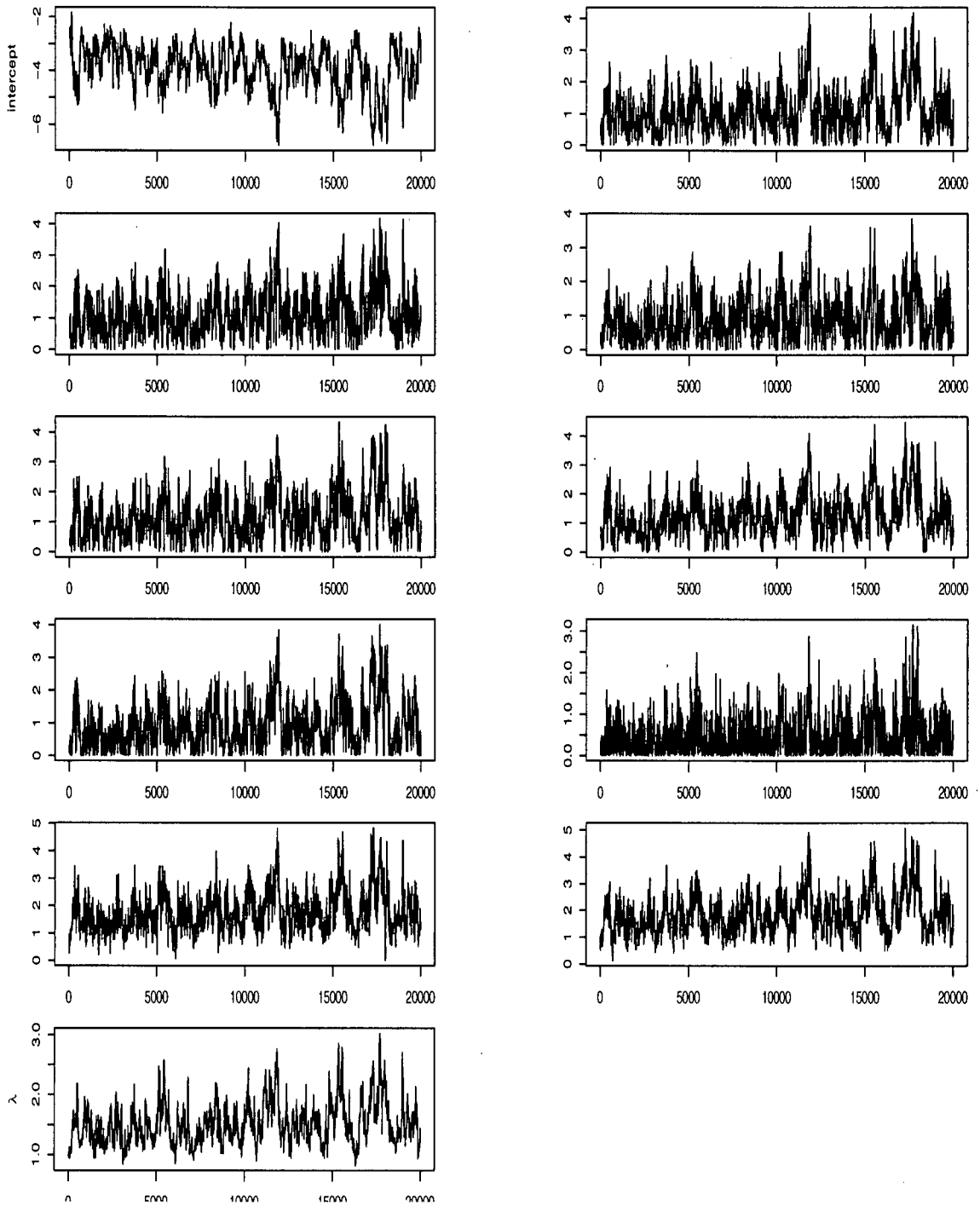

Figure 3.1: Sample path of the MCMC output where the first panel is for the intercept, panels 2-10 are of the coefficients and the last one is for λ.

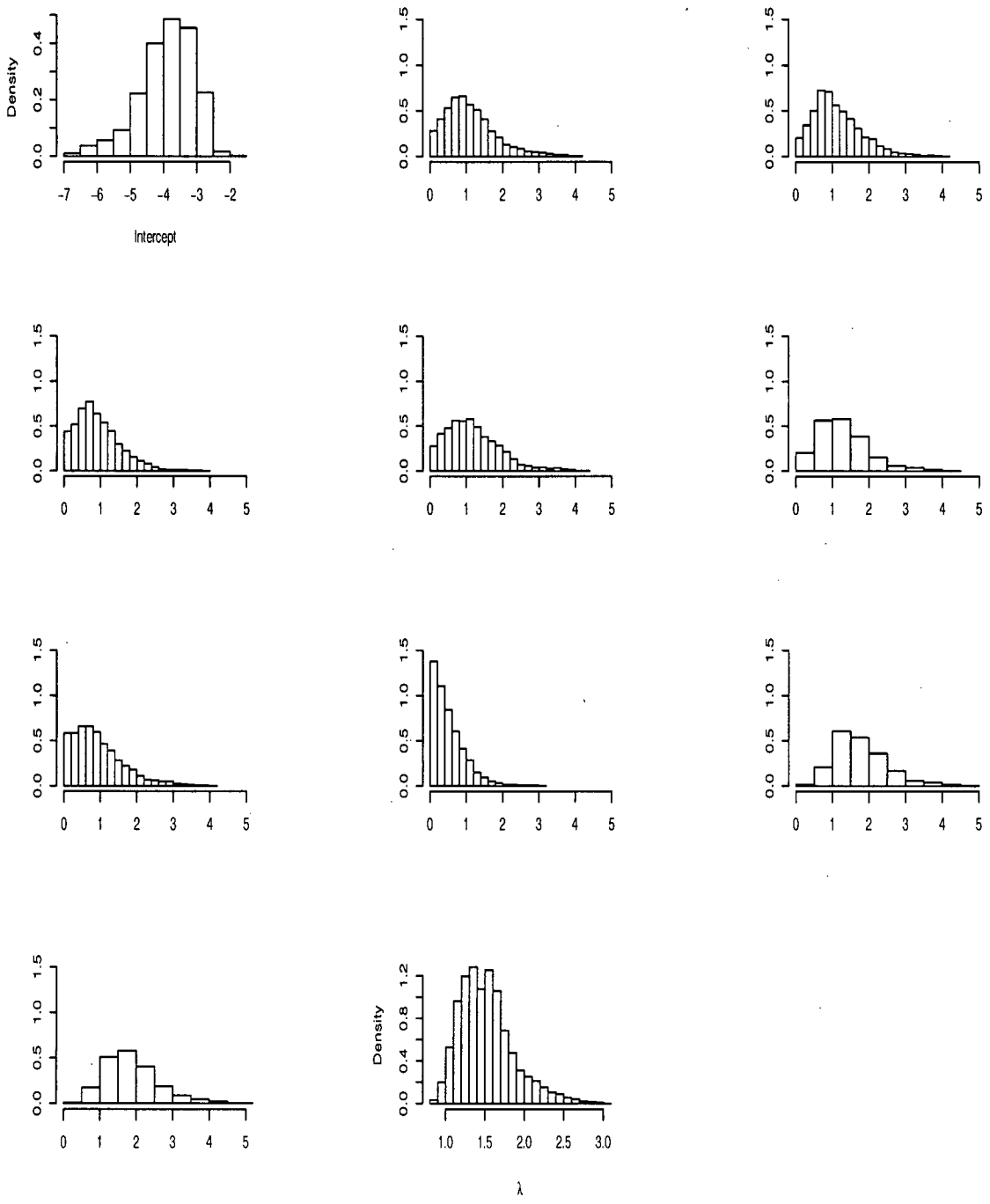

Figure 3.2: Posterior distribution of the parameters. The first panel is for the intercept, panels 2-10 are of the coefficients and the last one is for λ.

support value of λ greater than one. The uncertainty is reflected by the larger posterior standard deviations. The reason that the estimated standard errors from the ordinary logistic regression fit are smaller is that these are obtained assuming $\lambda = 1$. But the posterior standard deviations are averaged over all possible values of λ. Moreover, correlation between β's and λ may give rise to larger posterior standard deviations. Consequently, as λ gets larger than one we end up with more spread-out posterior distributions of the parameters.

Again, when we fit ordinary logistic regression we assume that the effect of a particular risk factor is the same for all persons in the study, so it measures average effect of that risk factor irrespective of the levels of other risk factors. With the NAD model, on the other hand, when interpreting β_j as the effect of the j-th risk factor we consider a person for whom all other risk factors are absent. Moreover, for $\lambda > 1$ the effect of a particular risk factor decreases as the number of risk factors increases. Hence the NAD model gives larger estimated β's than does the ordinary logistic regression fit. As for example, the posterior mean $\hat{\beta}_9 = 1.8662$ represents the risk in logit of having FamHist as a potential risk factor assuming that the remaining risk factors X_1 to X_8 are absent. To get comparable estimates as the ordinary logistic regression model we compute the average effect of the risk factors by the following quantity using sample MCMC output of the parameters:

$$\begin{aligned} \text{average effect of } X_j \;=\; & \frac{1}{n}\sum_{i=1}^{n}\big[\text{logit}(X_j = 1, X_{(j)} = x_{i(j)}, \beta, \lambda) - \\ & \qquad \text{logit}(X_j = 0, X_{(j)} = x_{i(j)}, \beta, \lambda)\big], \end{aligned} \tag{3.1}$$

where $j = 1, 2, \ldots, 9$.

Equation (3.1) gives the average effect of a particular risk factor with respect to the empirical distributions of other predictors, from the difference in logits when that risk factor is from its lower level to higher level in the presence of the remaining risk factors. Column 1 in Table 3.4 gives the posterior mean of (3.1), denoted by AV_1. Column 2 was obtained by inserting posterior means of the parameters into Equation 3.1, denoted by AV_2.

From Table 3.4 we observe that AV_2 values are very close to those estimates obtained by fitting the ordinary logistic regression. The AV_1 values are slightly larger than the ordinary logistic regression estimates. Therefore, it can be concluded that both models indicate similar average effects of the risk factors, as can be seen from Figure 3.3 in which AV_2 values are plotted

Table 3.4: Estimated average effects of the nine predictors using the NAD model

Coefficients	AV_1	AV_2
X_1	0.46	0.45
X_2	0.49	0.46
X_3	0.38	0.35
X_4	0.51	0.50
X_5	0.58	0.57
X_6	0.39	0.38
X_7	0.17	0.14
X_8	0.91	0.90
X_9	0.98	0.98

against logistic regression estimates.

While estimating effects, ordinary logistic regression model assumes that the effect of a particular risk factor is the same for all persons. If the effects of X_j in (3.1) for person i varies slightly with i, the NAD model is very close to the ordinary logistic regression model. However, histograms of individual level effects which are plotted in Figure 3.4, show considerable variation of effects from person to person. We thus qualitatively conclude that the fitted NAD model is substantially different from the fitted ordinary logistic regression model.

3.1.1 Model Comparison

To compare the predictive power of the NAD model with that of the step-wise and ordinary logistic regression model (which is fit considering full model irrespective of the statistical significance of the risk factors), we use the cross-validation procedure as described in Chapter 1, Section 1.5. For purpose of cross-validation we randomly divide the whole data set into five approximately equal-sized segments to compute the predicted probabilities for these three models.

The sample MCMC output via the NAD model was obtained after iterating the Hybrid (HY) algorithm 20,000 times using $k = 4, \epsilon = 0.06, c = 0.5 * \log(2)$ and $\delta = 0.90$. The step-wise models have been selected by using the procedure described in Chapter 1, Section 1.4.1. For selecting variables we use the entry probability $p_{\mathrm{E}} = 0.20$ and the probability of removal is $p_{\mathrm{R}} = 0.25$. The predicted probabilities of the presence of the disease outcome are then computed

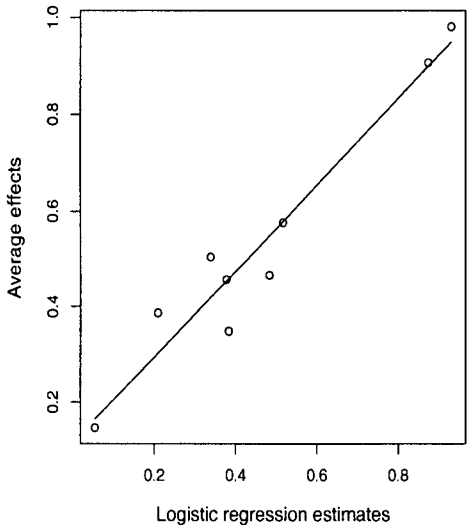

Figure 3.3: Scatterplot for comparing average effects with the logistic regression estimates

by:

$$Pr(Y = 1|X) = \frac{1}{1 + \exp(-g(X, \hat{\theta}))}\ , \tag{3.2}$$

where $g(X, \hat{\theta})$ is the corresponding estimated logit function for whichever model is under consideration. We compute Equation (3.2) for the NAD model by using the sample MCMC output and is based on the following relationship:

$$\Pr(\text{new person has disease} \mid \text{data}) = E\left\{\Pr(\text{new person has disease} \mid \beta, \lambda) \mid \text{data}\right\}.$$

The right hand side of the above relationship is approximated by:

$$E[h(\beta, \lambda) \mid \text{data}] \approx \frac{1}{m} \sum_{i=1}^{m} h(\beta^{(i)}, \lambda^{(i)}), \tag{3.3}$$

where h is given by Equation (3.2). As an alternative, we also computed (3.2) by plugging in the posterior means of the parameters. The computed predicted probabilities for the three models are presented in Figure 3.5 by boxplots.

We see from this figure that the NAD model predicts the presence of the disease outcome slightly better than the other two models. Its predictive performance about the absence of

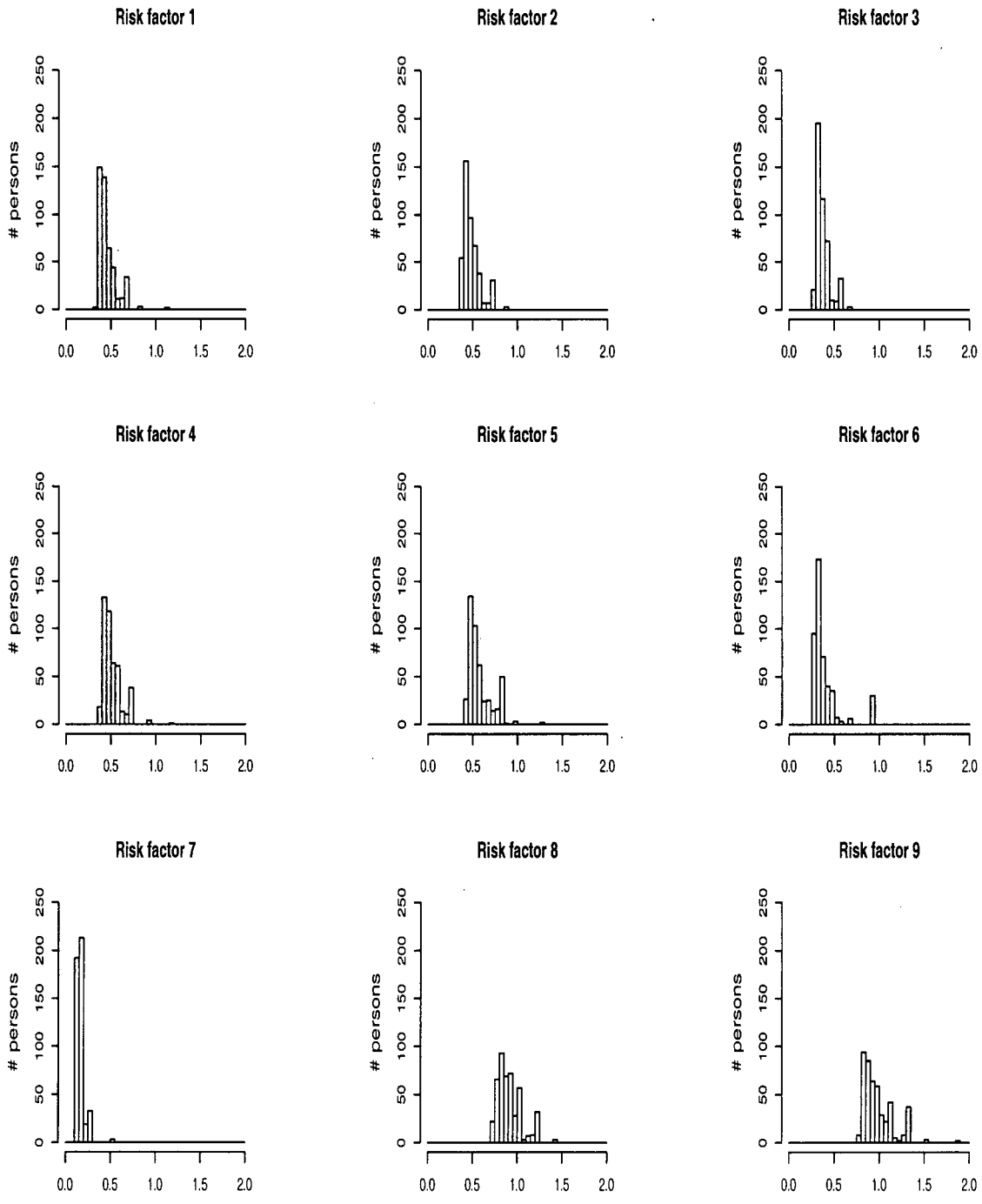

Figure 3.4: Distribution of individual level effects for all risk factors

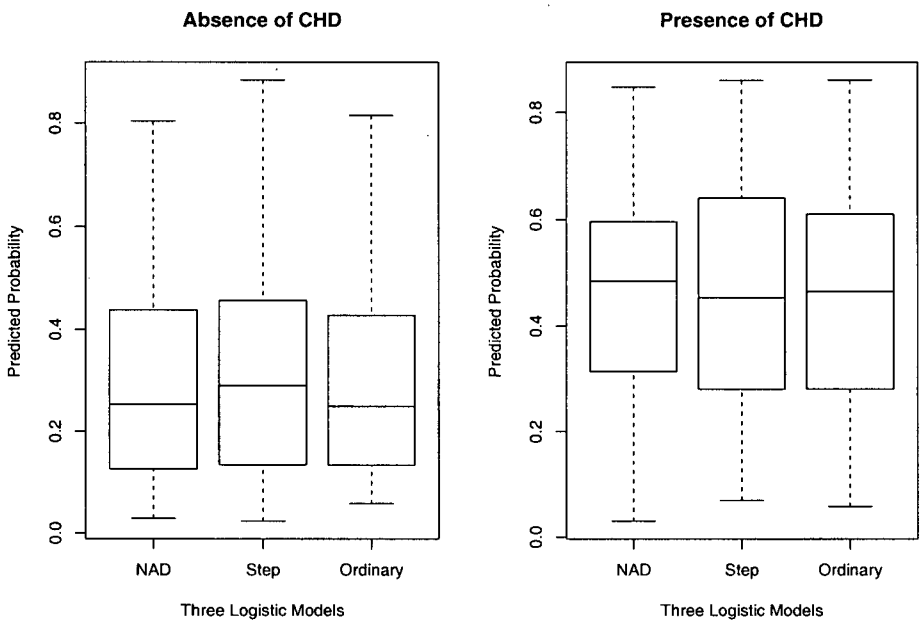

Figure 3.5: Boxplot for comparing performance of three logistic models

disease is better than that of the step-wise but similar to that of the ordinary. Note that, step-wise and the ordinary logistic regression models have the same abilities to predict the presence of disease outcome.

To have an intuitive idea of how similar is the NAD model with the ordinary logistic regression model we compute the fitted probabilities using both models from Equation (3.2). We use sample MCMC output to compute fitted probabilities by the NAD model. These are plotted against the fitted probabilities from the ordinary model in Figure 3.6.

From this figure we see that the fitted probabilities from both models are highly collinear and lie approximately on the straight line. Qualitatively, we can say that though both models indicate similar fitted probabilities, they bear different interpretations and getting a posterior mean of $\lambda > 1$ with the NAD model indicates presence of different kind of interactions.

The log-likelihoods of the predicted probabilities of the three models can be computed

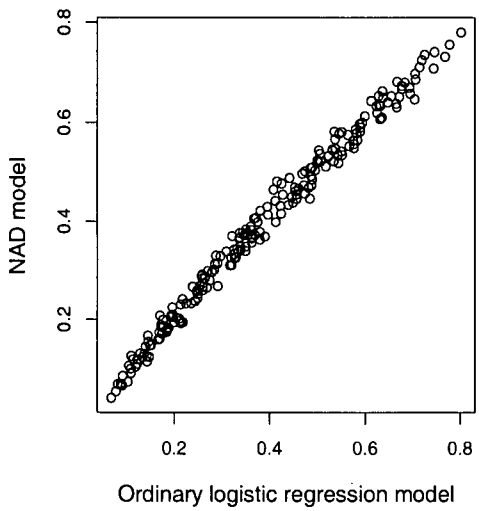

Figure 3.6: Scatterplot of fitted probabilities

by:

$$\begin{aligned} \text{Likelihood } L &= \prod_{i=1}^{n} \hat{p}_i^{y_i}[1-\hat{p}_i]^{(1-y_i)} \\ \log L &= \sum_{i=1}^{n} [y_i \log \hat{p}_i + (1-y_i)\log(1-\hat{p}_i)] \\ &= \sum_{i=1}^{n} \left[y_i \log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) + \log(1-\hat{p}_i) \right]. \end{aligned} \tag{3.4}$$

Using Equation (3.4) we have,

$$\begin{aligned} \log L(NAD) &= -270.0520, \\ \log L(NAD) &= -272.2717 \quad \text{(using posterior means)}, \\ \log L(step) &= -284.7087, \\ \log L(ordinary) &= -271.1916. \end{aligned}$$

Thus, the NAD model gives better log likelihood than the other two. To find out how much better is the predictive performance of the NAD model than that of the step-wise model, we

compute:

$$\exp\left[\text{average}\ \left\{\log\frac{L(NAD)}{L(step)}\right\}\right] = \exp\left[\frac{\log L(NAD) - \log L(step)}{n}\right] \tag{3.5}$$
$$= \exp\left(\frac{14.6567}{460}\right)$$
$$= 1.0324.$$

Hence the NAD model is only 3.24% better in predicting disease outcome than that of the step-wise model for this example.

In the next section we will investigate the second example to estimate the parameters and compare the performance of the three models following the same procedures as this example.

3.2 Example 2 : Scottish Heart Health Study (SHHS)

The Cardiovascular Epidemiology Unit of the University of Dundee, Scotland undertook a range of epidemiological studies in the mid 80's in order to understand the factors associated with CHD prevalence across Scotland. The Scottish Heart Health Study (SHHS) was one of these epidemiological studies with the objective to establish the levels of CHD risk factors in a cross-sectional sample of Scottish men and women aged 40-59 years drawn from different localities. This particular dataset contains a sample of 4049 males with six risk factors from the huge study. Of the 4049 subjects 196 had the CHD event. The response variable CHD is coded as

$$\text{CHD} = \begin{cases} 1 & \text{if the individual had a CHD event} \\ 0 & \text{if not} \end{cases}$$

Description of the six risk factors are given in Table 3.5.

Table 3.5: Code sheet for the Scottish Heart Health Study

#	Description	Codes/Values	Name
1	Age	years	AGE
2	Total Cholesterol	mmol/l	TOTCHOL
3	Body Mass Index (weight/height^2)	kg/m^2	BMI
4	Systolic Blood Pressure	mmHg	SYSTOL
5	Smoking status	1 = never smoked, 2 = ex-smoker, 3 = current smoker	SMK
6	Activity in leisure	1 = active, 2 = average, 3 = inactive	ACTIVITY

The first four quantitative risk factors are converted into binary variables by thresholding in comparison to their means. For converting smoking status we assume 'never smoked' category as low level and 'ex-smoker' & 'current smoker' categories combinedly as high level. For activity in leisure we assume 'active' & 'average' as low level and 'inactive'as high level.

We now draw samples from (2.18) by implementing hybrid algorithm. We assume $k = 5$, $c = \frac{1}{2}\log(2)$, $\epsilon = 0.035$, and $\delta = 0.90$. The initial values for β's are the estimates from the ordinary logistic regression fit, and $\phi^0 = 0$. We iterate the algorithm for 20,000 times and the acceptance rate is 90%. The sample plots of the MCMC output are given in Figure 3.7. Figure 3.8 shows the histogram of the posterior distribution. From Figure 3.7 we see, though

Table 3.6: Summary results of the posterior distribution from SHHS

	Posterior		90% CI	
Variable	Means	Std Dev.	5%	95%
Intercept	-5.56	0.87	-7.23	-4.41
X_1	0.87	0.61	0.11	1.95
X_2	1.81	0.76	0.86	3.21
X_3	0.54	0.54	0.01	1.55
X_4	1.68	0.74	0.72	2.98
X_5	1.57	0.82	0.48	3.10
X_6	0.63	0.58	0.02	1.75
λ	1.75	0.47	1.13	2.58

Table 3.7: Summary results from the ordinary logistic regression fit of the SHHS

			95% CI	
Coefficients	Estimate	Std Error	Lower	Upper
Intercept	-4.37	0.26	-4.88	-3.85
X_1	0.22	0.15	-0.08	0.51
X_2	0.75	0.17	0.45	1.06
X_3	0.09	0.15	-0.20	0.38
X_4	0.68	0.15	0.37	0.98
X_5	0.54	0.21	0.12	0.95
X_6	0.08	0.19	-0.30	0.46

the mixing of the sampler output is slow for some of the parameters, the MC in each case is stationary. The posterior means and standard deviations, computed using the 20,000 iterations are given in Table 3.6. The 90% equal-tailed credible intervals (CI) are also presented in Table 3.6, obtained from 5% and 95% sample quantiles. The risk factors are listed in the same order as given in Table 3.5. The summary results of the ordinary logistic regression fit is given in Table 3.7.

As with Example 1, the restriction imposed on the coefficients to be positive has some effect on some of the posterior distributions of the parameters as can be seen from Figure 3.8.

Comparing the estimates from Table 3.7 with the posterior means and standard deviations in Table 3.6 we see the same features as before, that is, posterior means and standard deviations are larger than the estimates and standard errors of the ordinary logistic regression

fit as we expect. The comparable estimates that can be obtained by (3.1) from the sample MCMC output are given in Table 3.8.

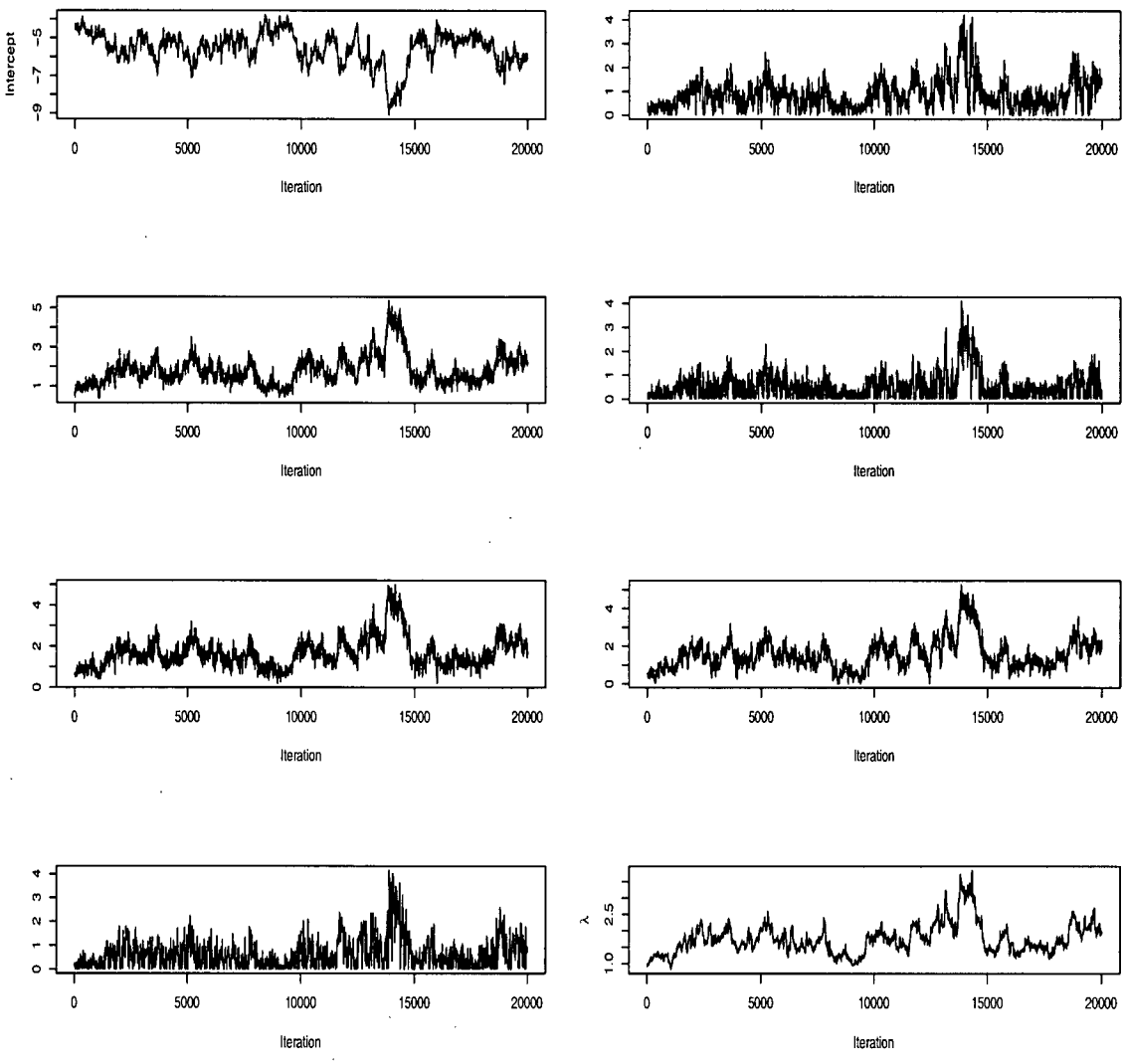

Figure 3.7: Sample path of the MCMC output where the first panel is for the intercept, panels 2-8 are of the coefficients and the last one is for λ.

AV_1 values are the posterior means of (3.1) obtained by using every tenth sample of the MCMC output. AV_2 values are computed from (3.1) by inserting the posterior means of the parameters. Table 3.8 reflects the same fact that the AV_2 values are very close to the estimates of the ordinary logistic regression fit. Moreover, AV_1 and AV_2 values are not very different.

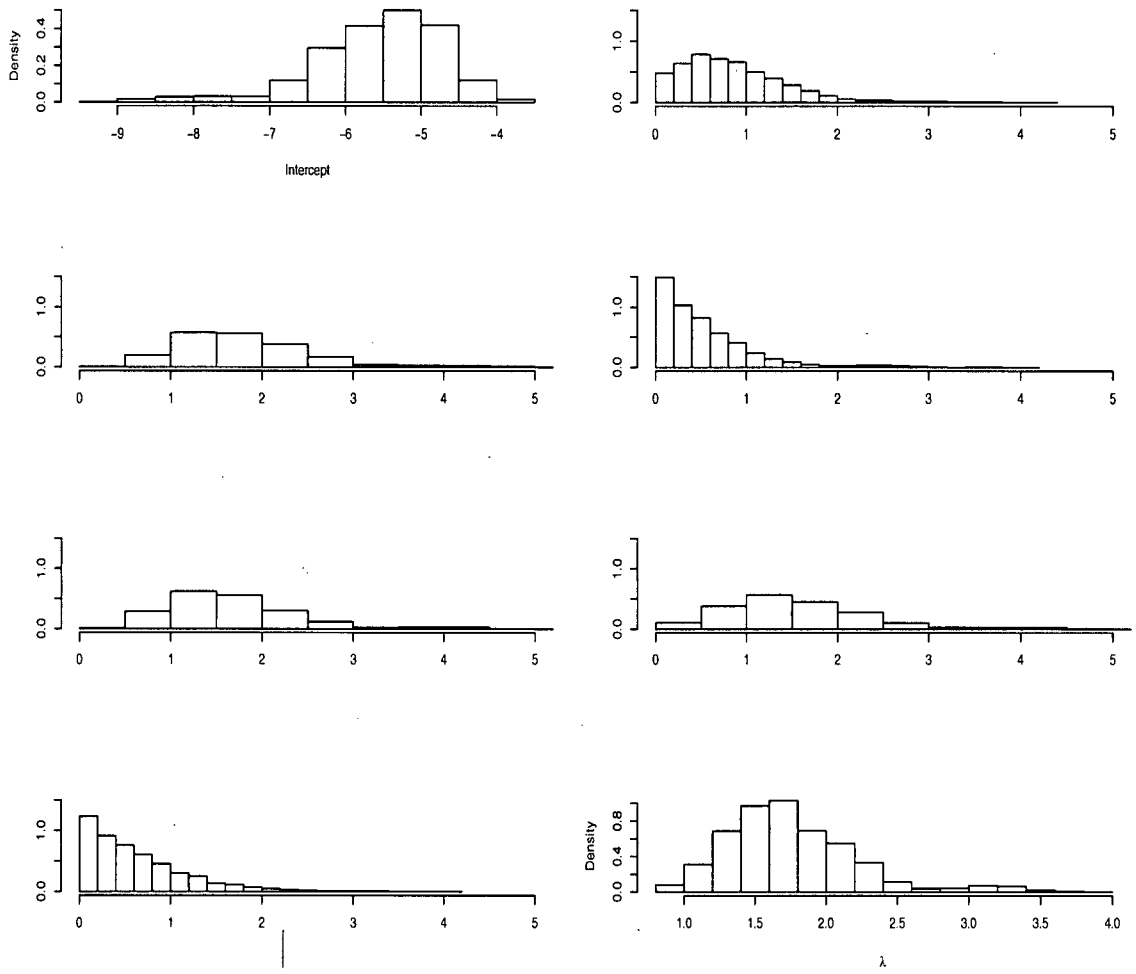

Figure 3.8: Posterior distribution of the parameters. The first panel is for the intercept, panels 2-8 are of the coefficients and the last one is for λ.

Therefore we can draw the same conclusion that both models indicate similar average effects of the risk factors as can be seen from Figure 3.9 in which AV_2 values are plotted against the estimates from the ordinary logistic regression fit.

Figure 3.10 shows the trace plots of the average effects of the six risk factors. Observing the trace plots and comparing with the sample paths of the parameters from Figure 3.7, we see that mixing of the sampler in these plots is better and looks good. Because of the high correlation between coefficients and λ, sampling from the NAD model was very hard using the MCMC methods. As a consequence, mixing of the sampler in Figure 3.7 is slow. But while

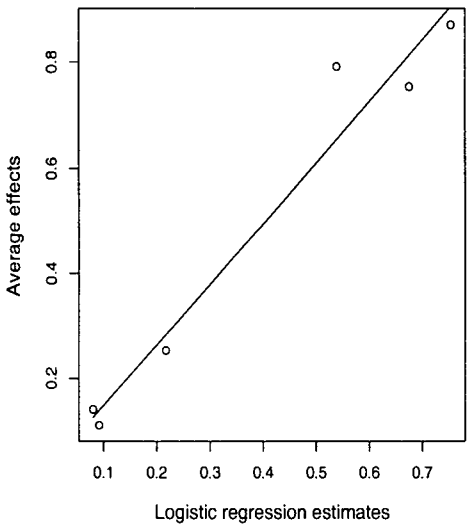

Figure 3.9: Scatterplot for comparing average effects with the logistic regression estimates

computing the average effects from the sample MCMC output, correlation has less effect and we get quite satisfactory mixing of the sample average effects.

Table 3.8: Estimated average effects of the six predictors from SHHS

Coefficients	AV_1	AV_2
X_1	0.26	0.25
X_2	0.86	0.87
X_3	0.13	0.11
X_4	0.75	0.75
X_5	0.79	0.79
X_6	0.16	0.14

3.2.1 Model Comparison

The predictive power of the NAD model will be compared with that of the step-wise and ordinary logistic regression model. As before we are using cross-validation procedure. The predicted probabilities were computed by (3.2). For the NAD model we follow the same procedure as

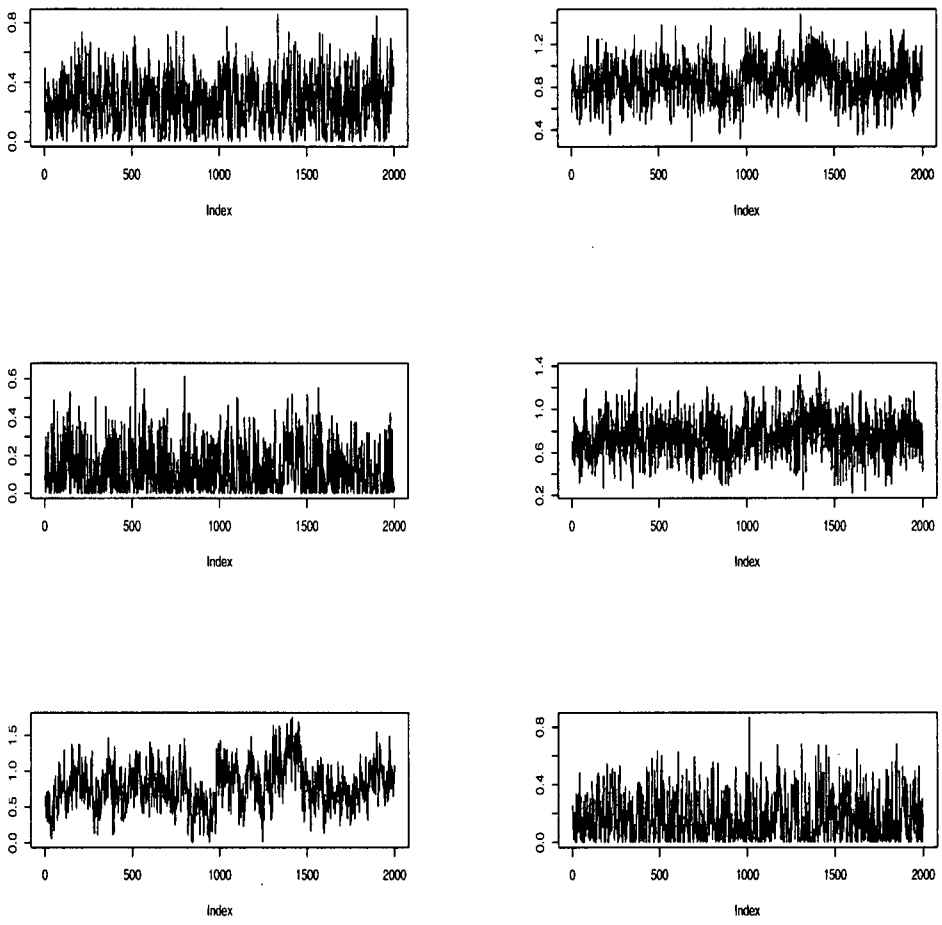

Figure 3.10: Trace plots of the average effects of the six predictors considering every tenth sample from the MCMC output of the parameters

Example 1 to compute the posterior means of the predicted probabilities using sample MCMC output. We also compute predicted probabilities plugging in the posterior means of the parameters. The log likelihoods computed using both procedure were presented. The predicted probabilities are presented in Figure 3.11. From the boxplots we see that the NAD model predicts the presence of CHD slightly better than step-wise and the ordinary logistic regression models. It also predicts the absence of CHD slightly better than the step-wise but the same as the ordinary.

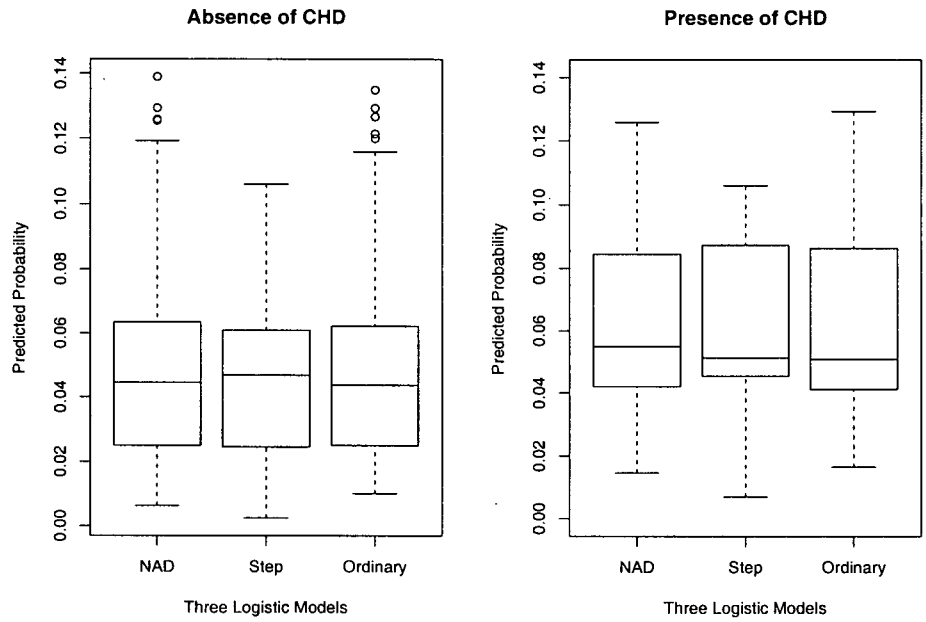

Figure 3.11: Boxplot for comparing performance of the three models

The log likelihoods for three models computed using (3.4) are given below:

$$\begin{aligned} \log L(NAD) &= -760.8353, \\ \log L(NAD) &= -761.2751 \quad \text{(using posterior means)}, \\ \log L(step) &= -765.9517, \\ \log L(ordinary) &= -763.1630. \end{aligned}$$

The model with the largest log likelihood is the best. Comparing the above three log predictive likelihoods we see that the NAD model has the largest log likelihood, but the differences with the other two are very small in light of the large sample size. Although we have a large sample in this data set the number of events is relatively very small (only 5% of the total cases), which might limit the predictive performance of the NAD model. For this reason we could not see a huge improvement. Using Equation (3.5) we see that the NAD model is only 0.13% better than the step-wise model. Considering these issues we can say that the NAD model has slightly better predictive power than step-wise for this example.

3.3 Example 3: Mystery data

The data for this example is from a real epidemiological study. The dataset contains 10,000 cases and twelve risk factors. Of the 10,000 subjects 1433 had the outcome event. The continuous risk factors are converted into binary variables. Due to reasons of data security and confidentiality, a description of the study and the risk factors is unavailable. The response variable y is defined as

$$y = \begin{cases} 1 & \text{if the outcome event is present,} \\ 0 & \text{if it is absent.} \end{cases}$$

The twelve predictors are denoted by $X_1, X_2, \ldots, X_{12}$. Figure 3.12 shows the distribution of the number of risk factors. There are 109 people in this study who do not have any risk factors and nobody has all twelve predictors. By observing the histogram we see that very few people have ten or eleven risk factors. The majority of the subjects have three or four risk factors.

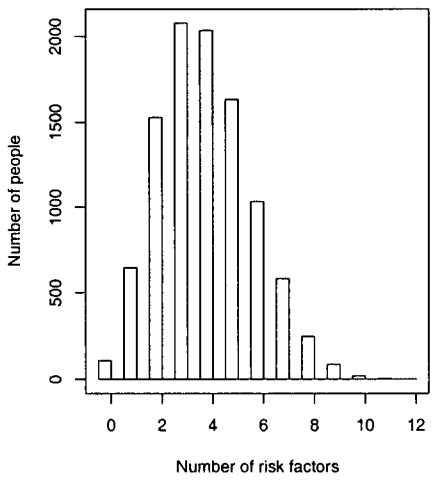

Figure 3.12: Distribution of the number of risk factors

To evaluate the Hybrid (HY) algorithm we assume $k = 4$, $c = \frac{1}{2}\log(2)$, $\epsilon = 0.016$, $\delta = 0.90$. The initial values for β's are the estimates from the ordinary logistic regression fit and $\phi^0 = 0$. We iterate the algorithm 20,000 times and the acceptance rate is 88%.

Figure 3.13 shows the sample path of the MCMC output. Figure 3.14 shows the histogram of the posterior distribution. From Figure 3.13 we see that it took some iterations for

Table 3.9: Summary results of the posterior distribution from Mystery data

	Posterior		90% CI	
Variable	Means	Std Dev.	5%	95%
Intercept	-4.11	0.28	-4.66	-3.75
X_1	0.87	0.23	0.56	1.31
X_2	0.50	0.22	0.18	0.92
X_3	1.19	0.23	0.85	1.61
X_4	0.28	0.20	0.02	0.61
X_5	1.12	0.24	0.81	1.57
X_6	0.72	0.26	0.32	1.18
X_7	1.47	0.25	1.11	1.94
X_8	1.32	0.26	0.94	1.81
X_9	0.34	0.20	0.04	0.69
X_{10}	0.56	0.28	0.10	1.03
X_{11}	0.26	0.19	0.02	0.64
X_{12}	0.71	0.30	0.21	1.23
λ	1.59	0.17	1.33	1.90

Table 3.10: Summary results from the ordinary logistic regression fit

			95% CI	
Coefficients	Estimate	Std Error	Lower	Upper
Intercept	-3.353	0.110	-3.568	-3.138
X_1	0.365	0.084	0.201	0.528
X_2	0.101	0.061	-0.019	0.221
X_3	0.494	0.061	0.374	0.614
X_4	0.029	0.081	-0.131	0.188
X_5	0.529	0.079	0.373	0.684
X_6	0.236	0.084	0.072	0.399
X_7	0.694	0.070	0.558	0.831
X_8	0.557	0.068	0.424	0.690
X_9	0.078	0.070	-0.058	0.214
X_{10}	0.157	0.100	-0.039	0.353
X_{11}	0.001	0.065	-0.125	0.128
X_{12}	0.265	0.111	0.047	0.483

Table 3.11: Estimated average effects of the twelve predictors from Mystery data

Variable	AV_1	AV_2
X_1	0.37	0.37
X_2	0.15	0.15
X_3	0.56	0.56
X_4	0.07	0.06
X_5	0.54	0.54
X_6	0.26	0.25
X_7	0.75	0.75
X_8	0.63	0.63
X_9	0.09	0.08
X_{10}	0.18	0.17
X_{11}	0.06	0.05
X_{12}	0.26	0.25

the chains to be stabilized. So we throw away first 5000 iterations as burn-in and compute posterior means from the remaining samples. The posterior means, standard deviations and 90% equal-tailed credible intervals, computed from 5% and 95% quantiles, are given in Table 3.9.

From Figure 3.14 we see the effect of the restriction imposed on the coefficients to be positive on some of the posterior distributions of the parameters.

Table 3.10 represents the summary results from the ordinary logistic regression fit. Comparing the estimates from Table 3.10 with those from Table 3.9, we again observe that the NAD model provides larger estimated β_i's and larger posterior standard deviations.

Comparable estimates, obtained by using (3.1), are given in Table 3.11. AV_1 values are obtained considering every tenth sample from the 15,000 sample MCMC output. AV_2 values are computed by plugging in the posterior means of the parameters into (3.1). Both AV_1 and AV_2 values are very close to the estimates from the ordinary logistic regression fit. Thus we can draw the same conclusion that both models indicate similar average effects of the risk factors, as can be seen from Figure 3.15 in which AV_2 values are plotted against logistic regression estimates.

Figure 3.16 shows the trace plots of the average effects of the twelve predictors. Comparing these plots with those from Figure 3.13 we observe that mixing is slow. Again, high correlation between λ and coefficients makes it difficult to sample even after several tuning the

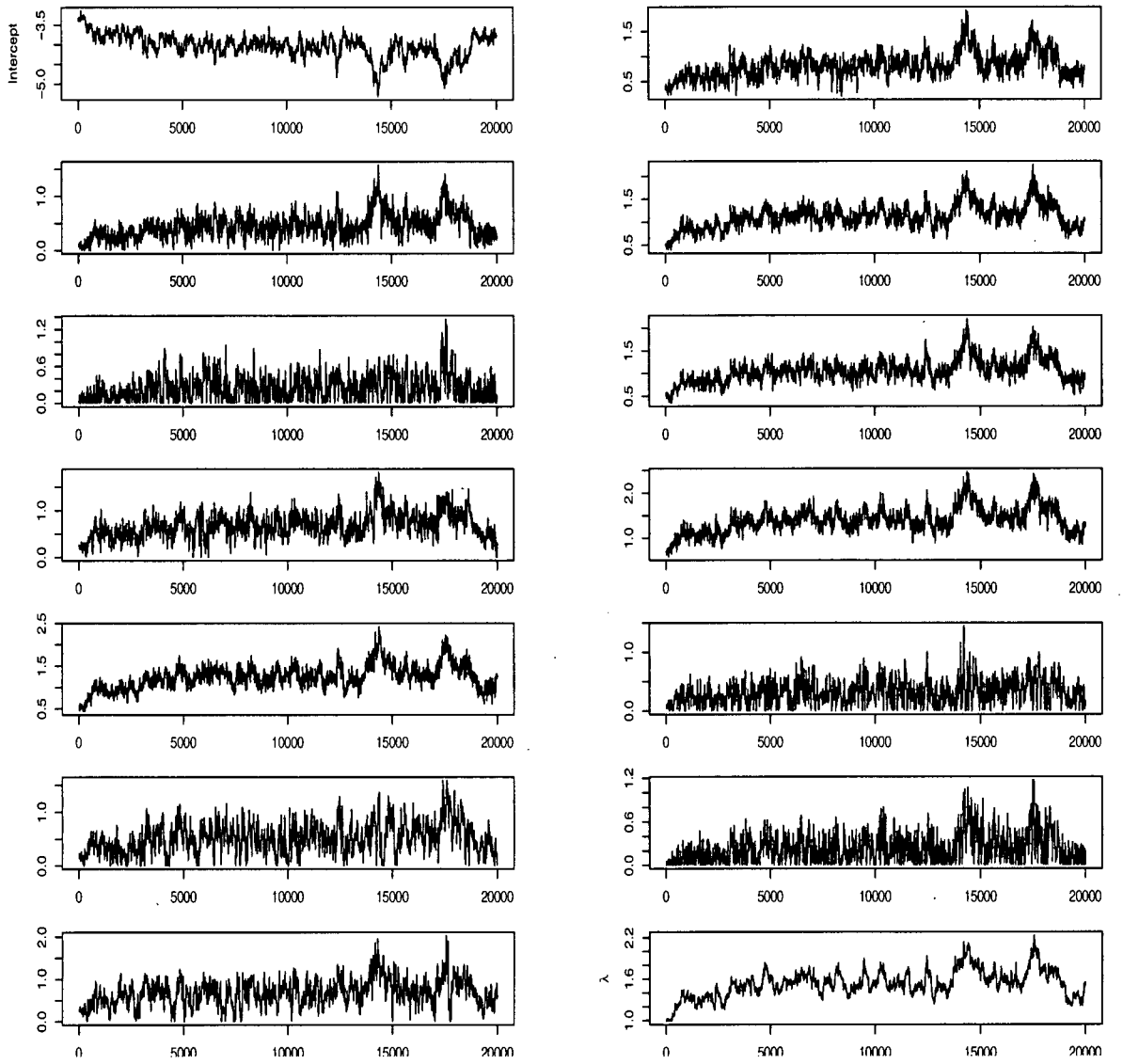

Figure 3.13: Sample path of the MCMC output where the first panel is for the intercept, panels 2-13 are of the coefficients and the last one is for λ

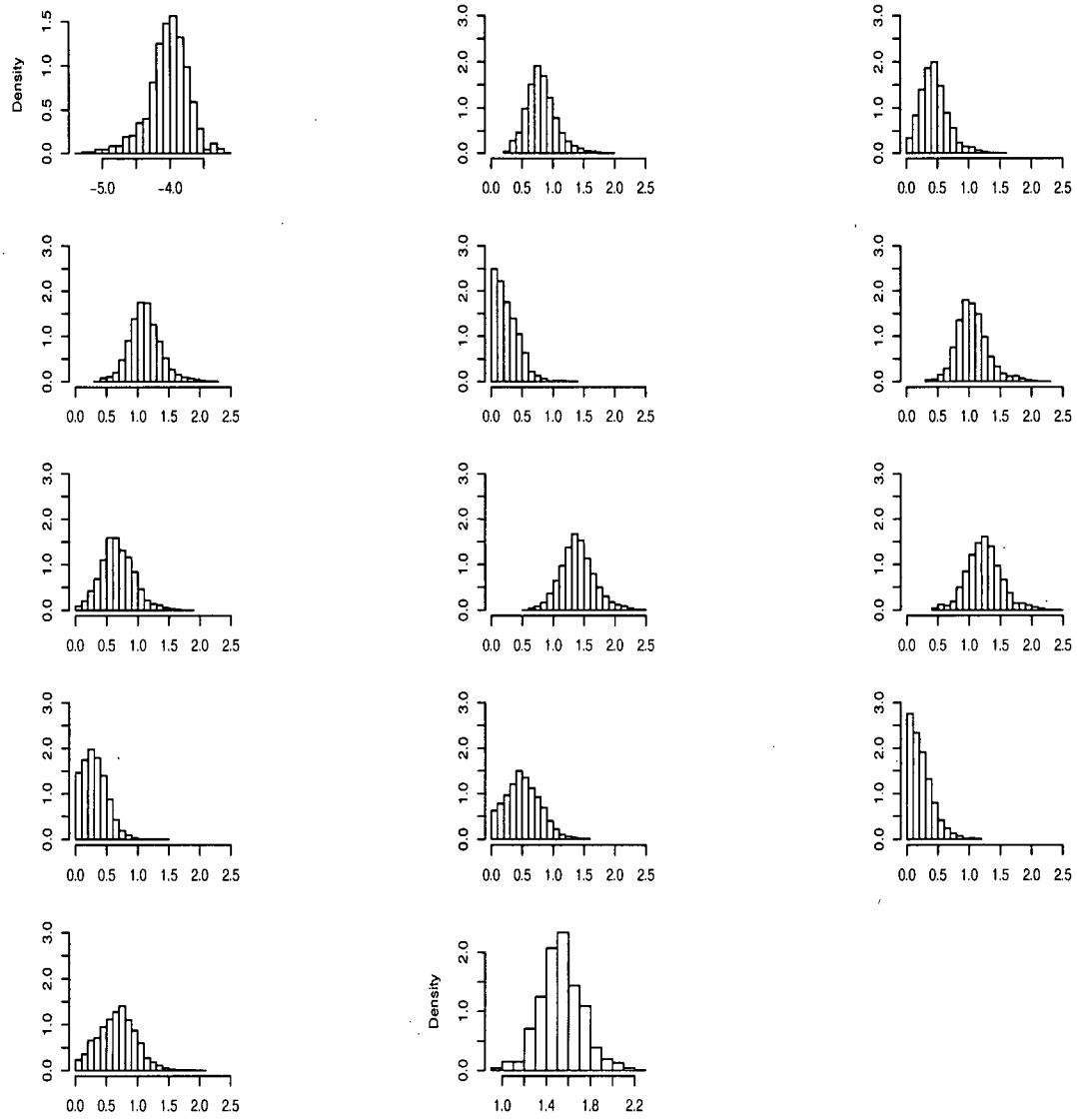

Figure 3.14: Posterior distribution of the parameters. The first panel is for the intercept, panels 2-13 are of the coefficients and the last one is for λ

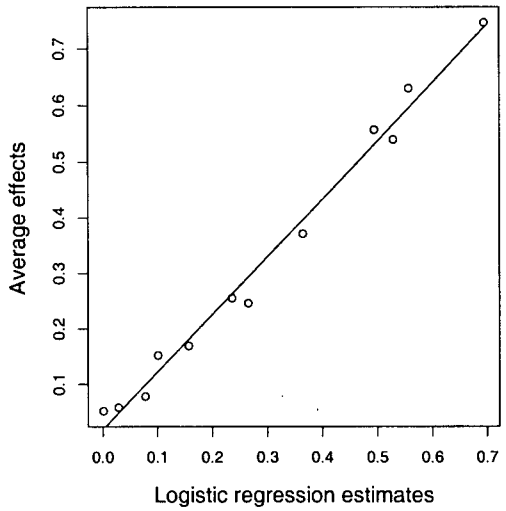

Figure 3.15: Scatterplot for comparing average effects with the logistic regression estimates

step size, and get proper mixing of the sampler output. As correlation has less effect on the average effects we are getting better mixing of the sample average effects.

3.3.1 Model Comparison

As with the Examples 1 and 2, we compare the predictive power of the NAD model with that of the step-wise and the ordinary logistic regression model. The cross-validation procedure is used. For selecting the step-wise model we use the probability for selecting a variable is 0.20 and for removing a variable is 0.25. The predicted probabilities were computed by (3.2). For implementing the HY algorithm with the NAD model we use $k = 4$, $\epsilon = 0.018$. The posterior means of the predicted probabilities were computed using sample MCMC output. Predicted probabilities by plugging in the posterior means of the parameters were also computed. The log predictive likelihoods were presented below. The predicted probabilities were plotted in Figure 3.17.

From the boxplots we observe that the NAD model predicts the presence and absence of the outcome the same as that of the step-wise model for this dataset. In both of the cases

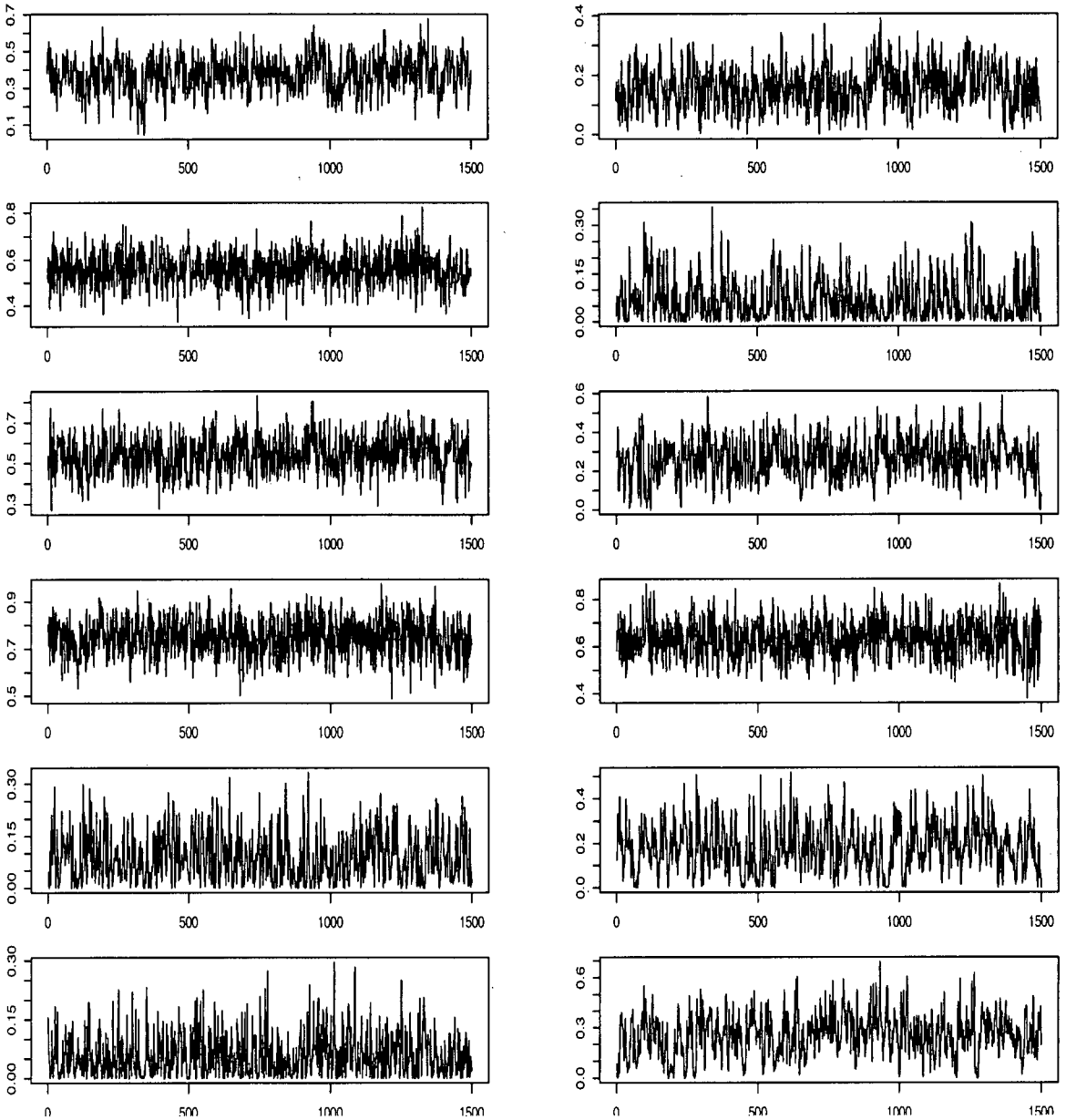

Figure 3.16: Trace plots of the average effects of the twelve predictors considering every tenth sample from the MCMC output of the parameters

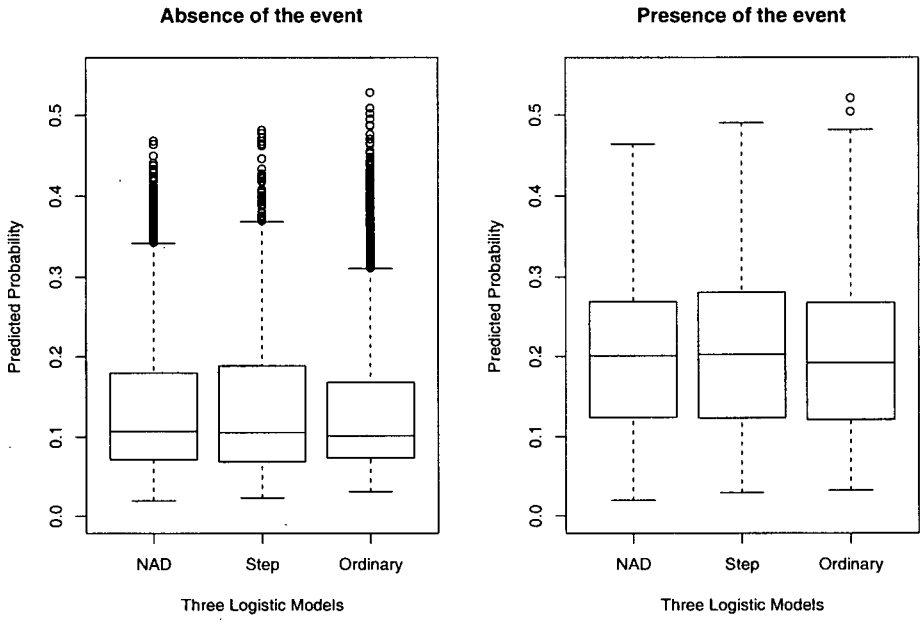

Figure 3.17: Boxplot for comparing performance of the three models

ordinary logistic regression model is slightly worse.

The log predictive likelihoods for three models computed by using (3.4) are as follows:

$$\begin{aligned} \log L(NAD) &= -3767.815, \\ \log L(NAD) &= -3768.121 \quad \text{(using posterior means)}, \\ \log L(step) &= -3779.009, \\ \log L(ordinary) &= -3772.059. \end{aligned}$$

Observing the above table we see that the NAD model has the largest log predictive likelihood, but the differences with the other two are very small. Though we have a large sample size and 15% cases have the event in this dataset, due to high correlation between λ and coefficients we could not improve the predicted probabilities. Nevertheless, using (3.5) we see that the NAD model has 0.11% better predictive power than the step-wise model.

Chapter 4

Discussion and Future Work

The most important finding to mention from the analysis of the data is that in each example we got a posterior mean of λ which is greater than one. This supports our belief about existence of possible kind of interactions other than the pairwise and our interpretations about the effects of the risk factors on the logit (Chapter 2, Section 2.2). From our experience we agree with Linde et. al. [13] that in course of analyzing the data, selection and inclusion of pairwise interaction terms into the step-wise logistic regression model is time-consuming and clumsy. We showed that the NAD model is slightly better than the step-wise and the ordinary logistic regression models in predicting the outcome of the response variable. We estimated average effects of the risk factors via the NAD model which are similar to the estimates of the ordinary logistic regression model and bear the equivalent interpretation. It is also worth mentioning that the ordinary logistic regression model is almost as capable as the step-wise in prediction.

For fitting the NAD model we assume binary covariates and restrict the sign of the coefficients to be positive. Due to the structural form of the NAD model we can not have negative covariates or negative values of the coefficients. Negative covariates can be included in the model by rescaling the minimum value to be zero. One useful technique that most epidemiologists apply when analyzing their data is to categorize the quantitative risk factors. However, if we have continuous risk factors scaled so that zero is the lowest risk and one is the highest risk, we can analyze them and interpret as usual conditional on $\beta_i > 0$. The estimated β or the posterior mean of a parameter can be interpreted as representing the risk difference (in logit scale) of a highest risk subject compared to a lowest risk subject when assuming the

remaining risk factors are absent. In all of the three examples used to fit the NAD model we converted the quantitative risk factors into binary variables and we coded the risk factors to match our *a priori* beliefs about the direction of effects, as reflected by the restriction to nonnegative coefficients. Indeed, in practical epidemiological context the question might arise: could we make full use of the NAD model with these restrictions? Or should we think of a new functional form? In spite of these limitations the positive aspect is that we can still analyze quantitative risk factors using the NAD model as long as we consider positive values of the coefficients. In many epidemiological studies we could guess the direction of the effect of many of the risk factors in advance. By knowing this fact we could choose the appropriate coding and interpret the coefficient estimate accordingly. However, in reality some predictor's characteristics in an epidemiological study may not be guessed completely in advance and if we were asked to analyze that predictor what we will do?

Suppose that all of the β_i's are positive where $i = 1, 2, \ldots, p$. Let us introduce the new parameters $\gamma_1, \gamma_2, \ldots, \gamma_p$. These parameters will govern the direction of the predictors and validate the structure of the NAD model by defining each of them as:

$$\gamma_i = \begin{cases} 1 & \text{if } X_i = 1 \text{ is the high risk level,} \\ 0 & \text{if } X_i = 0 \text{ is the high risk level.} \end{cases}$$

We have to estimate these parameters along with the other parameters already in the model. After introducing γ_i's the NAD model can be rewritten as:

$$\begin{aligned} logit\ Pr(Y = 1|X) &= \beta_0 + \Big\{\beta_1^{\lambda} X_1^{\gamma_1}(1 - X_1)^{1-\gamma_1} + \beta_2^{\lambda} X_2^{\gamma_2}(1 - X_2)^{1-\gamma_2} + \\ &\quad \ldots + \beta_p^{\lambda} X_p^{\gamma_p}(1 - X_p)^{1-\gamma_p}\Big\}^{1/\lambda}, \end{aligned} \tag{4.1}$$

where X_i's are binary variables or if continuous, scaled to be zero or one. This functional form in the logit scale can overcome the restriction of presuming to know the direction of each effect in advance.

Model (4.1) can be fit by following the Bayesian approach to model fitting discussed in Chapter 2, Section 2.3. This model now contains some discrete and some continuous parameters. Two procedures can be followed to evaluate the parameters, either (i) by fixing γ's and simulate β's and ϕ, or (ii) simulate all of them simultaneously. If γ's are known in advance, we go back to the NAD model and can follow the same procedures. If not, a prior distribution for γ,

say $\pi(\gamma)$, is need to be assumed along with β and ϕ. See George and McCulloch [6] for a discussion of assigning a prior distribution to γ when there are mixtures of discrete and continuous parameters. We have the similar situation here. The joint posterior distribution is then written as:

$$\pi(\beta, \phi, \gamma|y) \propto L(\beta, \phi, \gamma)\pi(\beta)\pi(\phi)\pi(\gamma).$$

Although analytical simplification of $\pi(\beta, \phi, \gamma|y)$ is intractable, existing MCMC methods such as the Gibbs sampler or Metropolis-Hastings algorithms (see Smith and Roberts [18], Chib and Greenberg [2] for an overview) can be used to explore the posterior $\pi(\gamma|y)$. Applied to the complete posterior $\pi(\beta, \phi, \gamma|y)$, such methods simulate a Markov chain

$$\beta^{(1)}, \phi^{(1)}, \gamma^{(1)}, \beta^{(2)}, \phi^{(2)}, \gamma^{(2)}, \ldots,$$

which converges in distribution to $\pi(\beta, \phi, \gamma|y)$. The embedded subsequence $\gamma^{(1)}, \gamma^{(2)}, \ldots$, thus converges to $\gamma \sim \pi(\gamma|y)$. Refer to George and McCulloch [6] again for additional comments on the simulation process. While we have not implemented this in the present thesis, it should be straightforward extension of the methods described.

Bibliography

[1] Agresti, Alan (1990). *Categorical Data Analysis.* John Wiley & Sons, New York.

[2] Chib, Siddartha and Greenberg, Edward (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician* Vol. 49, No. 4, 327-335.

[3] Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters* B 195, 216-222.

[4] Gamerman, Dani (1997). *Markov Chain Monte Carlo: Stochastic simulation for Bayesian inference.* Chapman & Hall, London.

[5] Gelman, A., Carlin, John B., Stern, Hal S. and Rubin, Donald B. (1995). *Bayesian Data Analysis.* Chapman & Hall, London.

[6] George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statist. Sinica* 7, 339-373.

[7] Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (1998). *Markov Chain Monte Carlo in Practice.* Chapman & Hall/CRC, London.

[8] Gustafson, Paul (1998). A guided walk Metropolis algorithm, *Statistics and Computing* 8, 357-364.

[9] Gustafson, P., MacNab, Y. C. and Wen, S. (2002). On the value of derivative evaluations and random walk suppression in Markov Chain Monte Carlo algorithms. Submitted.

[10] Hastie, T. J. and Tibshirani, R. J. (2001). *The Elements of Statistical Learning.* Springer, New York.

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

[12] Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression.* John Wiley & Sons, Inc. New York.

[13] Linde, Angelika van der and Osius, Gerhard (2001). Estimation of non-parametric multivariate risk functions in matched case-control studies with application to the assessment of interactions of risk factors in the study of cancer. *Statistics in Medicine* 20, 1639-1662.

[14] McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models,* Second edition. Chapman & Hall, London.

[15] 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.

[16] Neal, R. M. (1993). Bayesian learning via stochastic dynamics, in C.L. Giles, S.J. Hanson, and J.D. Cowan (eds.), *Advances in Neural Information Processing Systems* 5, pp. 475-482, San Mateo, California: Morgan Kaufmann.

[17] Rousseauw, J., du Plessis, J., Benade, A., Jordaan, P., Kotze, J., Jooste, P. and Ferreira, J. (1983). Coronary risk factor screening in three rural communities. *South African Medical Journal* 64, 430-436.

[18] Smith, A. F. M. and Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B* 55, 3-24.

[19] Woodward, Mark (1999). *Epidemiology: Study Design and Data Analysis.* Chapman & Hall/CRC. London.