# Model Selection for Generalized Linear Models via GLIB, with Application to Epidemiology [1]

Adrian E. Raftery
University of Washington

Sylvia Richardson
INSERM

December 21, 1993

[1]This is the first draft of a chapter for *Bayesian Biostatistics*, edited by Donald A. Berry and Darlene K. Strangl. Adrian E. Raftery is Professor of Statistics and Sociology, Department of Statistics, GN-22, University of Washington, Seattle, WA 98195, USA. Sylvia Richardson is Directeur de Recherche, INSERM Unité 170, 16 avenue Paul Vaillant Couturier, 94807 Villejuif CEDEX, France. Raftery's research was supported by ONR contract no. N-00014-91-J-1074, by the Ministère de la Recherche et de l'Espace, Paris, by the Université de Paris VI, and by INRIA, Rocquencourt, France. Raftery thanks the latter two institutions, Paul Deheuvels and Gilles Celeux for hearty hospitality during his Paris sabbatical in which part of this chapter was written. The authors are grateful to Christine Montfort for excellent research assistance and to Mariette Gerber, Michel Chavance and David Madigan for helpful discussions.

**Abstract**

Epidemiological studies for assessing risk factors often use logistic regression, log-linear models, or other generalized linear models. They involve many decisions, including the choice and coding of risk factors and control variables. It is common practice to select independent variables using a series of significance tests and to choose the way variables are coded somewhat arbitrarily. The overall properties of such a procedure are not well understood, and conditioning on a single model ignores model uncertainty, leading to underestimation of uncertainty about quantities of interest (QUOIs).

We describe a Bayesian modeling strategy that formalizes the model selection process and propagates model uncertainty through to inference about QUOIs. Each possible combination of modeling decisions defines a different model, and the models are compared using Bayes factors. Inference about a QUOI is based on an average of its posterior distributions under the individual models, weighted by their posterior model probabilities; the models included in the average are selected by the Occam's Window algorithm. In an initial exploratory phase, ACE is used to suggest ways to code the variables, but the final coding decisions are based on Bayes factors. The methods can be implemented using GLIB, an S function available free of charge from StatLib. We apply our strategy to an epidemiological study of fat and alcohol consumption as risk factors for breast cancer.

# Contents



# List of Tables

# 1 Introduction

Are fat and alcohol consumption associated with breast cancer? This question has been much discussed in the past ten years. In this chapter we introduce a new Bayesian modeling strategy for generalized linear models, and use it to reanalyze a case-control study in which fat and alcohol consumption and eight other risk factors were measured for 854 women (Richardson *et al.*, 1989). It is usual to analyze such studies by logistic regression (e.g. Breslow and Day, 1980). The analyst must make several decisions, including

*How to code the risk factors?* In epidemiological studies, risk factors are often coded in several categories (typically 2 to 5). This has the advantage of allowing easy calculation of relative risks. If the risk factor is not inherently categorical, however, it is somewhat arbitrary and different codings (e.g. a different number of categories, different breakpoints, or coding as a continuous variable) may yield different results.

*Which risk factors to include?* This is the key question in the study and is analogous to variable selection in regression. It is complicated by the fact that the data at hand do not support the inclusion of some of the "classical" risk factors, for which the evidence comes from previous studies.

Each possible combination of decisions defines a different statistical model for the data, so the number of models initially considered is very large. Even if one considers only two possible ways of coding each risk factor, and the possibility of omitting it, there are $3^{10}$, or about 60,000 possible models. Typically, the analyst selects one of those models by some combination of more or less arbitrary choices (for coding) and significance tests (for risk factor selection). He or she then makes inference about quantities of interest (QUOIs) conditionally on the selected model.

There are three main difficulties with such a strategy. First, the way in which the variables are coded is often somewhat arbitrary, and different choices might well give different results. Second, the basis for risk factor selection strategies based on sequences of significance tests is weak because (a) the properties of the overall strategy (as distinct from the individual tests) are not well understood, and (b) the final choice often comes down to a comparison of non-nested models, which cannot easily be done by significance tests. Third, by conditioning on a single model selected out of a large number, the uncertainty associ-

ated with the model selection process itself is ignored, and as a result the uncertainty about QUOIs is underestimated. This can lead, for example, to decisions that are riskier than one thinks (Hodges, 1987).

In this chapter we describe a Bayesian modeling strategy that avoids these difficulties. It consists of calculating the posterior probabilities of all the models considered, and averaging over them when making inferences about QUOIs, thereby taking account of model uncertainty. Several approximations are suggested so as to make computation and communication of the results easier. In an initial exploratory phase, the ACE algorithm is used to suggest a set of candidate codings for the risk factors. Instead of averaging over all possible models, we average over a reduced set of models given by the Occam's Window algorithm (Madigan and Raftery, 1994). The result is a practical and formally justifiable modeling strategy that avoids arbitrariness in variable coding and risk factor selection, and accounts for model uncertainty. It can be implemented using the GLIB software, available free of charge from StatLib.

In Section 2 we review the Bayesian approach to model selection and model uncertainty using Bayes factors, in the context of generalized linear models. In Section 3 we outline our modeling strategy, and in Section 4 we describe its application to the nutrition and breast cancer study. In Section 5 we give some information about the GLIB software.

# 2 Bayes Factors and Model Uncertainty

In this section, we review the basic ideas of Bayes factors and their use in accounting for model uncertainty. Sections 2.1–2.5 summarize material in Raftery (1988, 1993) and Kass and Raftery (1994), to which the reader is referred for more information.

## 2.1 Basic Ideas

We begin with data $D$ assumed to have arisen under one of the two models $M_1$ and $M_2$ according to a probability density $\mathrm{pr}(D|M_1)$ or $\mathrm{pr}(D|M_2)$. Given prior probabilities $\mathrm{pr}(M_1)$ and $\mathrm{pr}(M_2) = 1-\mathrm{pr}(M_1)$, the data produce posterior probabilities $\mathrm{pr}(M_1|D)$ and $\mathrm{pr}(M_2|D) = 1 - \mathrm{pr}(M_1|D)$. Since any prior opinion gets transformed to a posterior opinion through consideration of the data, the transformation itself represents the evidence provided by the data. In fact, the same transformation is used to obtain the posterior probability, regardless of the prior probability. Once we convert to the odds scale (odds = probability/(1

– probability)), the transformation takes a simple form. From Bayes' Theorem we obtain

$$\mathrm{pr}(M_k|D) = \frac{\mathrm{pr}(D|M_k)\mathrm{pr}(M_k)}{\mathrm{pr}(D|M_1)\mathrm{pr}(M_1) + \mathrm{pr}(D|M_2)\mathrm{pr}(M_2)} \quad (k = 1, 2),$$

so that

$$\frac{\mathrm{pr}(M_1|D)}{\mathrm{pr}(M_2|D)} = \frac{\mathrm{pr}(D|M_1)}{\mathrm{pr}(D|M_2)} \frac{\mathrm{pr}(M_1)}{\mathrm{pr}(M_2)},$$

and the transformation is simply multiplication by

$$B_{12} = \frac{\mathrm{pr}(D|M_1)}{\mathrm{pr}(D|M_2)}, \tag{1}$$

which is the *Bayes factor*. Thus, in words,

$$\text{posterior odds} = \text{Bayes factor} \times \text{prior odds},$$

and the Bayes factor is the ratio of the posterior odds of $M_1$ to its prior odds. When the models $M_1$ and $M_2$ are equally probable *a priori* so that $\mathrm{pr}(M_1) = \mathrm{pr}(M_2) = 0.5$, the Bayes factor is equal to the posterior odds in favor of $M_1$.

In the simplest case, when the two models are single distributions with no free parameters (the case of "simple versus simple" testing), $B_{12}$ is the likelihood ratio. In other cases, when there are unknown parameters under either or both of the models, the Bayes factor is still given by (1) and, in a sense, it continues to have the form of a likelihood ratio. Then, however, the densities $\mathrm{pr}(D|M_k)$ $(k = 1, 2)$ are obtained by *integrating* (not maximizing) over the parameter space, so that in equation (1),

$$\mathrm{pr}(D|M_k) = \int \mathrm{pr}(D|\theta_k, M_k)\mathrm{pr}(\theta_k|M_k)\mathrm{d}\theta_k, \tag{2}$$

where $\theta_k$ is the vector of parameters under $M_k$, $\mathrm{pr}(\theta_k|M_k)$ is its prior density and $\mathrm{pr}(D|\theta_k, M_k)$ is the probability density of $D$ given the value of $\theta_k$, or the likelihood function of $\theta$.

The quantity $\mathrm{pr}(D|M_k)$ given by equation (2) is the marginal probability of the data, since it is obtained by integrating the joint density of $(D, \theta_k)$ given $D$ over $\theta_k$. It is also sometimes called the marginal likelihood, or the integrated likelihood.

One important use of the Bayes factor is as a summary of the evidence for $M_1$ against $M_2$ provided by the data. It can be useful to consider twice the logarithm of the Bayes factor, which is on the same scale as the familiar deviance and likelihood ratio test statistics. We use the following rounded scale for interpreting $B_{12}$, which is based on that of Jeffreys (1961), but is more granular and slightly more conservative than his.

| $B_{12}$ | $2 \log B_{12}$ | Evidence for $M_1$ |
|---|---|---|
| $< 1$ | $< 0$ | Negative (supports $M_2$) |
| 1 to 3 | 0 to 2 | Not worth more than a bare mention |
| 3 to 12 | 2 to 5 | Positive |
| 12 to 150 | 5 to 10 | Strong |
| $> 150$ | $> 10$ | Very strong |

Kass and Raftery (1994) provide a full review of Bayes factors, to which readers are referred for more information. They discuss issues that we do not dwell on here, including whether it is valid or useful to test point hypotheses or select models at all, and the comparison between Bayes factors and non-Bayesian significance testing for nested models.

## 2.2 Accounting for Model Uncertainty

When more than two models are being considered, the Bayes factors yield posterior probabilities of all the models, as follows. Suppose that $(K+1)$ models, $M_0, M_1, \ldots, M_K$, are being considered. Each of $M_1, \ldots, M_K$ is compared in turn with $M_0$, yielding Bayes factors $B_{10}, \ldots, B_{K0}$. Then the posterior probability of $M_k$ is

$$\mathrm{pr}(M_k|D) = \alpha_k B_{k0} / \sum_{r=0}^{K} \alpha_r B_{r0}, \tag{3}$$

where $\alpha_k = \mathrm{pr}(M_k)/\mathrm{pr}(M_0)$ is the prior odds for $M_k$ against $M_0$ $(k = 0, \ldots, K)$; here $B_{00} = \alpha_0 = 1$. For discussion of the prior model probabilities, $\mathrm{pr}(M_k)$, see Section 2.5.

The posterior model probabilities given by equation (3) lead directly to solutions of the prediction, decision-making and inference problems that take account of model uncertainty. The posterior distribution of a QUOI, $\Delta$, such as a relative risk or the probability that someone who does not have the disease now will get it later, is

$$\mathrm{pr}(\Delta|D) = \sum_{k=0}^{K} \mathrm{pr}(\Delta|D, M_k)\mathrm{pr}(M_k|D), \tag{4}$$

where $\mathrm{pr}(\Delta|D, M_k) = \int \mathrm{pr}(\Delta|D, \theta_k, M_k)\mathrm{pr}(\theta_k|D, M_k)\mathrm{d}\theta_k$. In a certain sense, equation (4) is guaranteed to give better out-of-sample predictions on average than conditioning on any single model (Madigan and Raftery, 1994). The composite posterior mean and standard deviation (or point estimate and standard error) are given by

$$E[\Delta|D] = \sum_{k=0}^{K} \hat{\Delta}_k \mathrm{pr}(M_k|D), \tag{5}$$

$$\mathrm{SD}[\Delta|D]^2 = \mathrm{Var}[\Delta|D] = \sum_{k=0}^{K} \left(\mathrm{Var}[\Delta|D, M_k] + \hat{\Delta}_k^2\right) \mathrm{pr}(M_k|D) - E[\Delta|D]^2, \qquad (6)$$

where $\hat{\Delta}_k = E[\Delta|D, M_k]$ (Raftery, 1992).

This general approach has been used in several previous analyses of medical and epidemiological data. Racine *et al.* (1986) showed how this method may be used to make inference about a treatment effect in the presence of uncertainty about the existence of a carryover effect. Raftery (1993) showed that model uncertainty was large and was an important source of uncertainty about the relative risk in a classic study of oral contraceptive use as a risk factor for myorcardial infarction. He showed how it could be accounted for using this framework. Madigan and Raftery (1994) did similar analyses of coronary heart disease risk factors and the diagnosis of scrotal swellings. In their examples, they found that out-of-sample predictive performance was better if one took account of model uncertainty than if one conditioned on *any* single model that might reasonably have been selected.

The posterior probability that a given parameter is nonzero can be calculated from equation (4); it is the sum of the posterior probabilities of the models in which the parameter is present. In epidemiology this can be the posterior probability that a particular risk factor is associated with the disease; see Section 4.3 for an example.

The number of models, $K$, can be so large that it is not practical to compute equation (4) directly. Two ways of getting around this have been suggested. One of these, "Occam's Window", consists of averaging over a much smaller set of models, chosen by criteria based on standard norms of scientific investigation (Madigan and Raftery, 1994). We use this here and describe it in Section 3.1. The other, Markov chain Monte Carlo model composition ($\mathrm{MC}^3$), approximates equation (4) using a Markov chain that moves through model space (Madigan and York, 1993).

## 2.3 Approximating Bayes Factors by the Laplace Method

The Laplace method for integrals (e.g. de Bruijn, 1970, Section 4.4) is based on a Taylor series expansion of the real-valued function $f(u)$ of the $p$-dimensional vector $u$, and yields the approximation

$$\int e^{f(u)} \mathrm{d}u \approx (2\pi)^{p/2} |A|^{\frac{1}{2}} \exp\{f(u^*)\}, \qquad (7)$$

where $u^*$ is the value of $u$ at which $f$ attains its maximum, and $A$ is minus the inverse Hessian of $f$ evaluated at $u^*$. When applied to equation (2) it yields

$$p(D|M_k) \approx (2\pi)^{p_k/2} |\Psi|^{\frac{1}{2}} \mathrm{pr}(D|\tilde{\theta}_k, M_k) \mathrm{pr}(\tilde{\theta}_k|M_k), \qquad (8)$$

where $p_k$ is the dimension of $\theta_k$, $\tilde{\theta}_k$ is the posterior mode of $\theta_k$, and $\Psi_k$ is minus the inverse Hessian of $h(\theta_k) = \log\{\mathrm{pr}(D|\theta_k, M_k)\mathrm{pr}(\theta_k|M_k)\}$, evaluated at $\theta_k = \tilde{\theta}_k$. Arguments similar to those in the Appendix of Tierney and Kadane (1986) show that in regular statistical models the relative error in equation (8), and hence in the resulting approximation to $B_{10}$, is $O(n^{-1})$.

One can approximate the marginal likelihood $\mathrm{pr}(D|M_k)$ in any regular statistical model using equation (8), but this requires the posterior mode $\tilde{\theta}_k$ and minus the inverse Hessian $\Psi_k$, which are not routinely available. Software that calculates the maximum likelihood estimator (MLE) $\hat{\theta}_k$, the deviance or the likelihood ratio test statistic, and the observed or expected Fisher information matrix, $F_k$, or its inverse, $V_k$, is much more widespread, and includes, for example, GLIM. Here we describe an approximation proposed by Raftery (1993) that is based on equation (8) but uses only these widely available quantities.

Suppose that the prior distribution of $\theta_k$ is such that $E[\theta_k|M_k] = \omega_k$ and $\mathrm{Var}\,[\theta_k|M_k] = W_k$. Then approximating $\tilde{\theta}_k$ by a single Newton step starting from $\hat{\theta}_k$ and substituting the result into equation (8) yields the approximation

$$2\log B_{10} \approx \chi^2 + (E_1 - E_0). \tag{9}$$

In equation (9), $\chi^2 = 2\{\ell_1(\hat{\theta}_1) - \ell_0(\hat{\theta}_0)\}$, where $\ell_k(\hat{\theta}_k) = \log\left(\mathrm{pr}(D|\theta_k, M_k)\right)$ is the log-likelihood; $\chi^2$ is the standard likelihood-ratio test statistic when $M_0$ is nested within $M_1$. Also,

$$E_k = 2\lambda_k(\hat{\theta}_k) + \lambda_k'(\hat{\theta}_k)^T(F_k+G_k)^{-1}\{2 - F_k(F_k+G_k)^{-1}\}\lambda_k'(\hat{\theta}_k) - \log|F_k+G_k| + p_k\log(2\pi), \tag{10}$$

where $G_k = W_k^{-1}$, $\lambda_k(\theta_k) = \log\mathrm{pr}(\theta_k|M_k)$ is the log-prior density, and $\lambda_k'(\hat{\theta}_k)$ is the $p_k$-vector of derivatives of $\lambda_k(\theta_k)$ with respect to the elements of $\theta_k$ $(k = 0, 1)$.

This approximation is closer to the basic Laplace approximation (8) when $F_k$ is the observed than the expected Fisher information, and so one would expect it also to be generally more accurate in this case. Arguments similar to those of Kass and Vaidyanathan (1992) show that when $F_k$ is the observed Fisher information, the relative error is $O(n^{-1})$, while when $F_k$ is the expected Fisher information, the relative error increases to $O(n^{-\frac{1}{2}})$. When the prior is normal, equation (10) becomes

$$-E_k = (\hat{\theta}_k - \omega_k)^T C_k(\hat{\theta}_k - \omega_k) + \log|W_k| + \log|F_k + G_k|. \tag{11}$$

In equation (11), $C_k = G_k H_k F_k H_k G_k + (I - H_k G_k)^T G_k (I - H_k G_k)$, where $H_k = (F_k + G_k)^{-1}$.

## 2.4 Application to Generalized Linear Models

Raftery (1993) has shown how the Laplace approximation of Section 2.3 can be used to calculate Bayes factors for generalized linear models using only standard GLIM output, or equivalent.

Suppose that $y_i$ is the dependent variable, and that $x_i = (x_{i1}, ..., x_{ip})$ is the corresponding vector of independent variables, for $i = 1, ..., n$. The model $M_1$ is defined by specifying $\mathrm{pr}(y_i|x_i, \beta)$ in such a way that $E[y_i|x_i] = \mu_i$, $\mathrm{Var}[y_i|x_i] = \sigma^2 v(\mu_i)$, and $g(\mu_i) = x_i\beta$, where $\beta = (\beta_1, ..., \beta_p)^T$; here $g$ is called the link function. The $n \times p$ matrix with elements $x_{ij}$ is denoted by $X$, and it is assumed that $x_{i1} = 1 \quad (i = 1, ..., n)$. For the moment we assume that $\sigma^2$ is known; the case where $\sigma^2$ is unknown will be discussed later.

We now derive approximate expressions for the Bayes factor for the null model $M_0$, defined by setting $\beta_j = 0 \quad (j = 2, ..., p)$, against $M_1$. The likelihoods for $M_0$ and $M_1$ can be written down explicitly, and so, once the prior has been fully specified, the approximation (8) can be computed. However, this approximation is not easy to compute for generalized linear models using readily available software.

By contrast, when applied to generalized linear models, the approximation of equations (9) and (10) is easier to compute. It is an analytic non-iterative function of the MLE, the deviance and the Fisher information matrix, and so can be calculated directly from GLIM output or equivalent. If $DV_k$ is the deviance for $M_k$, then $\chi^2 = (DV_0 - DV_1)/\sigma^2$. The expected Fisher information matrix is $F_1 = \sigma^{-2} X^T W X$, where $W = \mathrm{diag}\{w_1, \ldots, w_n\}$, and $w_i^{-1} = g'(\hat{\mu}_i^{(1)})^2 v(\hat{\mu}_i^{(1)})$ (McCullagh and Nelder, 1989). Here $\hat{\mu}_i^{(1)} = g^{-1}(x_i\hat{\beta}^{(1)})$, where $\hat{\beta}^{(1)}$ is the MLE of $\beta$ conditional on $M_1$. Similarly, $F_0 = \sigma^{-2} n g'(\hat{\mu}_i^{(0)})^2 v(\hat{\mu}_i^{(0)})$, where $\hat{\mu}_i^{(0)} = g^{-1}(\hat{\beta}^{(0)})$, $\hat{\beta}^{(0)}$ being the MLE of $\beta_1$ conditional on $M_0$. The observed and expected Fisher information matrices are equal when $g$ is the canonical link function, and so the approximations are more accurate in this case. These values of $\chi^2$, $\hat{\beta}^{(0)}$, $\hat{\beta}^{(1)}$, $F_0$ and $F_1$ can be substituted directly into equations (9) and (10).

This solution can be extended to situations where the dispersion parameter, $\sigma^2$, is unknown, such as Poisson and binomial models with overdispersion, or models with normal or Gamma errors. One may proceed as before with $\sigma^2$ replaced by an estimate $\tilde{\sigma}^2$, as McCullagh and Nelder (1989) do for estimation. A reasonable estimate would be $\tilde{\sigma}^2 = P/(n-p)$, where $P$ is Pearson's goodness-of-fit statistic for the most complex model considered, as advocated by McCullagh and Nelder (1989, p. 91 and p. 127).

The method can be used to compare different link functions. Suppose that we are comparing two models $M_1$ and $M_2$, which have the same independent variables $X$ and variance

function $v$, but different link functions $g_1$ and $g_2$. Then the parameters $\beta^{(1)}$ and $\beta^{(2)}$ under the two models are on different scales and so should have different prior distributions. Thus we calculate $2\log B_{10}$ and $2\log B_{20}$ as before, but with different priors obtained separately for each link function, for example as in Section 2.5 below. We then compare $M_1$ and $M_2$ using the relation $2\log B_{21} = 2\log B_{20} - 2\log B_{10}$.

Finally, the method may be used to compare different error distributions. Consider the comparison of two models, $M_1$ and $M_2$ which have the same independent variables $X$ but different variance functions and/or different error distributions; they may also have different link functions. We can continue to use the same general framework because equation (8) still gives the marginal likelihood for each model, and Bayes factors and posterior model probabilities are then available from equations (1) and (3) as before.

The parameters $\beta^{(1)}$ and $\beta^{(2)}$ of the two models are on different scales and so should have different prior distributions. The first step is then to specify the prior distribution of $\beta$ for each model, as in Section 2.5. We may then use the same approximation as before. Equation (9) becomes

$$2\log B_{21} \approx \chi^{2*}_{21} + (E_2 - E_1). \tag{12}$$

In equation (12),

$$\chi^{2*}_{21} = (\sigma_1^{-2}\mathrm{DV}_1 - \sigma_2^{-2}\mathrm{DV}_2) + 2(\ell_2^{\mathrm{sat}} - \ell_1^{\mathrm{sat}}), \tag{13}$$

where $\ell_k^{\mathrm{sat}}$ is the maximal log-likelihood achievable with the link function and error distribution of model $M_k$ $(k = 1, 2)$; this will typically be the log-likelihood under the saturated model. In equation (13), $\sigma_k^2$ is the dispersion parameter for $M_k$, which is either known or estimated as above. In equation (12), $E_k$ is given as before by equation (10).

## 2.5 A Reference Set of Proper Priors

The prior distributions $\mathrm{pr}(\theta_k|H_k)$ $(k = 1, 2)$ are necessary. This may be considered both good and bad. Good, because it is a way of including other information about the values of the parameters. Bad, because these prior densities may be hard to set when there is no such information.

Raftery (1993) considered the situation where there is little prior information, and suggested a reasonable set of prior distributions for this situation. We now briefly review this proposal.

We first consider the case where $g(\mu) = \mu$ and $v(\mu) = 1$, and where the variables have been standardized to have mean 0 and variance 1. In this situation, we denote the parameters

by $\gamma$. We assume that the prior distribution of $(\gamma|M_1)$ is normal; in fact the results depend rather little on the precise functional form. We also assume that $(\gamma_2, \ldots, \gamma_p)$ are independent *a priori*; this corresponds to the situation where the individual variables are of interest in their own right, which is often implicit in the testing situation. We further assume that the prior is *objective* for the testing situation in the sense of Berger and Sellke (1987), that is, symmetric about the null value of $\gamma$, namely $(\gamma_1, 0, \ldots, 0)^T$, and non-increasing as one moves away from the null value.

These assumptions lead to the prior $(\gamma|M_1) \sim N(\nu, U)$, where $\nu = (\nu_1, 0, \ldots, 0)$ and $U = \text{diag}\{\psi^2, \phi^2, \ldots, \phi^2\}$. The prior for $\gamma$ under $M_0$ is just the conditional prior distribution of $\gamma$ under $M_1$ given that $\gamma_2 = \ldots = \gamma_p = 0$, namely $(\gamma_1|M_0) \sim N(\nu_1, \psi^2)$. This is transformed back into a prior on the original parameters $\beta$.

This is extended to generalized linear models with other link and variance functions by noting that then estimation is equivalent to weighted least squares with the adjusted dependent variable $z_i = g(\mu_i) + (y_i - \mu_i)g'(\hat{\mu}_i)$ and weights $w_i$ (McCullagh and Nelder, 1989). The prior is the same as before, but in the transformation back to the original $\beta$, $y$ is replaced by $z$ and all the summary statistics are weighted.

When several models are considered, it is desirable that the priors be consistent with each other in the sense that if $M_2$ is defined by setting restrictions $\rho(\beta) = 0$ on the parameters of $M_1$, then $\text{pr}(\beta|M_1) = \text{pr}\,(\beta|M_2, \rho(\beta) = 0)$. A reasonable way to ensure this is to obtain a prior for the largest model as above, and then derive the priors for other models by conditioning on the constraints that define them.

The prior distribution has three user-specified parameters: $\nu_1$, $\psi$ and $\phi$. Bayes factors tend to be insensitive to $\nu_1$ and $\psi$; see Kass and Raftery (1994) for an explanation based on Kass and Vaidyanathan (1992). Raftery (1993) found $\nu_1 = 1$ and $\psi = 1$ to be reasonable values. Bayes factors are more sensitive to $\phi$, however, with larger values of $\phi$ tending to favor simpler models. Raftery (1993) defined a reasonable range of values of $\phi$ by requiring that the prior not contribute much evidence in favor of either model, whether the models being compared are nested or not. These requirements are to some extent in conflict: for non-nested models it implies that $\phi$ be large, while for nested models it implies that $\phi$ not be too large. Balancing these two desiderata in a certain sense gives $\phi = e^{\frac{1}{2}} = 1.65$, and requiring that the priors not contribute evidence "worth more than a bare mention" beyond what is unavoidable leads to the range $1 \leq \phi \leq 5$. The resulting priors are then transformed back to the original scale for the variables; results for other choices of $g(\mu)$ and $v(\mu)$ are obtained by weighting the cases appropriately.

The result is a *reference set of proper priors* for generalized linear models. While they are mildly data-dependent, they do have properties that one would associate with genuine subjective data-independent priors that represent a small amount of prior information.

## 2.6 Prior Model Probabilities

When there is little prior information about the relative plausibility of the models considered, taking them all to be equally likely *a priori* is a reasonable "neutral" choice. When the number of models is small or moderate, this is intuitively appealing and understandable. When the number of models is very large, however, one could worry about whether this choice might have unintended perverse consequences, as has happened before with other "uniform" priors. In our experience with very large model spaces (up to $10^{12}$ models) involving several kinds of model and about 20 data sets, we have found no such perverse effects (Madigan *et al.*, 1993; Raftery, Madigan and Hoeting, 1993; Madigan and Raftery, 1994). In addition, we have found inference using Occam's Window (Section 3.1) to be quite robust to moderately large changes in prior model probabilities (e.g. halving or doubling the prior odds).

If it is available, prior information can easily be taken into account by adjusting the prior model probabilities. In variable selection problems, prior information often takes the form of prior evidence for the inclusion of a variable, rather than for an individual model. Suppose that this is the only kind of prior information available and that model $M_k$ is specified by a vector $(\delta_{k1}, \ldots, \delta_{kp})$, where $\delta_{ki} = 1$ if the $i$-th variable is included and 0 if not. Then, if $\pi_i$ is the prior probability that the $i$-th variable has an effect, and if the prior information about different variables is approximately independent, it is reasonable to specify

$$\mathrm{pr}(M_k) \propto \prod_{i=1}^{p} \left[ \pi_i^{\delta_{ki}} (1 - \pi_i)^{(1-\delta_{ki})} \right]. \tag{14}$$

In epidemiology, prior information often takes the form of strong evidence from previous studies for the inclusion of a particular risk factor. If the previous studies combined are more informative about the risk factor than the present one, then it is a reasonable approximation to set the corresponding $\pi_i = 1$ in equation (14), and hence to consider only models that include that risk factor. This approximation is reasonable in the sense that the true prior probability $\pi_i$ is so close to 1 that the posterior probability would also be close to 1, almost regardless of the data at hand. Thus setting $\pi_i = 1$ yields a posterior distribution of QUOIs close to what would result from using the true $\pi_i$. This provides a formal rationale for the common practice of "controlling" for particular independent variables even when the data at hand provide little evidence for their inclusion in the model.

In our experience with moderate to large data sets, we have found it sufficient to restrict attention to prior model probabilities of the form (14) with $\pi_i = 0, \frac{1}{2}$ or 1. This might not be enough, however, for small data sets such as small clinical trials, where more careful assessment and elicitation of prior model probabilities might be needed.

# 3 Modeling Strategy

We now describe our modeling strategy. We start by reviewing Occam's Window, which embodies our criteria for selecting a set of models. We then describe how we use ACE to select an initial candidate set of codings for each risk factor. Finally we give the iterative algorithm used for combining these two steps.

## 3.1 Occam's Window

This algorithm was proposed by Madigan and Raftery (1994) and selects a subset of the models initially considered. It involves averaging over a much smaller set of models than in (4), thereby making it easier to compute and to communicate model uncertainty.

They argued that if a model is far less likely *a posteriori* than the most likely model, then it has been discredited and should no longer be considered. Thus models not belonging to

$$\mathcal{A}' = \left\{ M_k : \frac{\mathrm{pr}(M_k|D)}{\max_l\{\mathrm{pr}(M_l|D)\}} \geq c \right\},$$

should be excluded from equation (4), where $c$ is a fairly small number ($\ll 1$) chosen by the data analyst. By analogy with the common 5% significance level used in frequentist tests, we have used $c = 0.05$. Appealing to Occam's razor, they also exclude from (4) any model which receives less support from the data than a simpler model that is nested within it, namely those belonging to

$$\mathcal{B} = \left\{ M_k : \exists M_l \in \mathcal{A}, M_l \subset M_k, \frac{\mathrm{pr}(M_l|D)}{\mathrm{pr}(M_k|D)} > 1 \right\}.$$

Then equation (4) is replaced by

$$\mathrm{pr}(\Delta|D) = \frac{\sum_{M_k \in \mathcal{A}} \mathrm{pr}(\Delta|M_k, D)\mathrm{pr}(D|M_k)\mathrm{pr}(M_k)}{\sum_{M_k \in \mathcal{A}} \mathrm{pr}(D|M_k)\mathrm{pr}(M_k)},$$

where $\mathcal{A} = \mathcal{A}' \backslash \mathcal{B}$.

Typically the number of terms in (4) is reduced to 25 or less, and often to as few as one or two, even when the number of models initially considered is very large. This procedure

mimics the evolutionary process of model selection which is typical of science. The final solution is fairly robust to the initial class considered, in the sense that most initial classes that contain $\mathcal{A}$ give the same result.

When the total number of models initially considered, $K$, is relatively small, the models in $\mathcal{A}$ can be found directly by considering each model in turn. When $K$ is very large, however, this may not be feasible, and Madigan and Raftery (1994) proposed a tree-based algorithm (the "Up-Down Algorithm") that is computationally feasible in this case. It has been successfully applied to situations where there are up to $10^{12}$ models.

## 3.2 Selecting Transformations with ACE

The ACE (Alternating Conditional Expectations) algorithm (Breiman and Friedman, 1985) is a way of finding nonparametric transformations of the independent and dependent variables in linear regression such that the relationship is as linear as possible. See DeVeaux (1989) for an exposition.

This can be extended to generalized linear models by noting that estimation there is equivalent to weighted least squares with the adjusted dependent variable $z_i$ and the weights $w_i$ given in Section 2.5. The $z_i$ and the $w_i$ involve the conditional expectations $\mu_i$, and the method can be implemented by initially estimating the $\mu_i$ from the full model with all risk factors, assuming linearity. For logistic regression, $z = \text{logit}(\mu) + (y - \mu)/\{\mu(1 - \mu)\}$ and $w = \{\mu(1 - \mu)\}^{-1}$, while for Poisson log-linear models, $z = \log \mu + (y - \mu)/\mu$ and $w = \mu^{-1}$ (dropping the subscript $i$'s). The transformation of the dependent variable must always be specified to be linear. We implemented ACE using the S-PLUS function `ace`.

We use the ACE output to *suggest* parametric transformations of the variables, if possible ones that are substantively interpretable; we do *not* use the non-parametric transformations estimated by ACE in their raw form. The transformations suggested often have the form of a threshold effect, with no change in the expected value of the dependent variable below or beyond a certain threshold value. This kind of transformation often fits the data better *and* is more interpretable than the commonly recommended addition of a quadratic term to deal with nonlinearity. Sometimes, the ACE output suggests that the variable be categorized and how this should be done. ACE has also been used in this way by Raftery, Lewis, Aghajanian and Kahn (1992) and Raftery, Lewis and Aghajanian (1993).

Often it is reasonable to assume that the relationship between disease and risk factor is monotonic, and imposing this constraint can be very helpful in obtaining useful output from ACE. We generally make this assumption unless the relationship is clearly non-monotonic.

To detect strong non-monotonicities, we run ACE a first time without the monotonicity constraints, and then a second time with them.

Generalized additive models provide an approach to modeling in the presence of non-linearity that is similar in concept to ACE and is conceived as a direct generalization of generalized linear models (Hastie and Tibshirani, 1990). However, the software to estimate them (in S-PLUS) does not yet allow for monotonicity constraints, and so we have found ACE more useful.

### 3.3 An Iterative Strategy

The methods we have described are linked in the following iterative algorithm:

1. Run ACE with the adjusted dependent variable $z_i$ and weights $w_i$, specifying the transformation of the dependent variable to be linear, but without monotonicity constraints. Identify independent variables whose estimated ACE transformations are strongly non-monotonic.

2. Run ACE again, this time with the monotonicity constraints for all the indepenent variables except those for which strongly non-monotonic relationships were identified in step 1.

3. Using the ACE output, identify candidate parametric transformations of the independent variables. Possibilities include power and log transformations, threshold and piecewise linear functions, categorical (i.e. piecewise constant) transformations, and the addition of quadratic and higher order polynomial terms.

4. Using the "preferred" codings of the independent variables from step 3, run GLIB using as models all relevant subsets of the independent variables.

5. For each of the most likely models *a posteriori* from step 4, calculate Bayes factors for it against perturbed versions of itself in which other candidate transformations are used, for each independent variable in turn. If any of the other transformations is preferred, use it and return to step 4.

6. Apply the rules of Section 3.1 to the models from step 4 so as to find the models in Occam's Window. Run GLIB again using only those models. Use the results to make inference about QUOIs or to report the main conclusions from the analysis and the remaining uncertainties.

We have found that ACE is highly effective at suggesting good transformations, and that the iteration between steps 4 and 5 is often not needed. If the number of models is too large, it will not be possible to execute step 4 directly, and it will have to be replaced by a version of the Up-Down algorithm of Madigan and Raftery (1994). Software to do this for generalized linear models has not yet been produced, but it can sometimes be done manually; see Raftery (1992) for an example.

# 4 An Epidemiological Application: Nutrition and Breast Cancer

This section will be concerned with illustrating the use of Bayes factor in developing a strategy for analysing an epidemiological study. We start by a brief description of the study.

## 4.1 Description of Study

The association of dietary factors and breast cancer (BC) has been widely discussed during the past decade and interest has been particularly focused on the role played by fat and alcohol intake. Epidemiological studies have given somewhat inconsistent results with respect to these latter risk factors, even though alcohol has been found to be positively associated with BC in a majority of studies (Howe *et al.*, 1991; Van der Brandt *et al.*, 1993).

The part played by dietary risk factors was tested in a case-control study which took place in Montpellier (France) between 1983 and 1987. A dietary history questionnaire administered by interview was used to measure the intake of total fat and its constituants, other nutrients and alcohol consumption. In the original analysis of this study, a positive association was found with alcohol (Richardson *et al.*, 1989) and fat intake (Richardson *et al.*, 1991), but the evidence for the role of fat intake was not overwhelming. Classical risk factors for BC, in particular those concerning the reproductive history, were also analysed (Ségala *et al.*, 1991). The reader is referred to the above mentioned references for a detailed description of the protocol and the analyses which were carried out. In the following, we shall focus our reanalysis of these data on evaluating the association with alcohol and fat intake. The sample consists of 854 women (379 cases and 475 controls) for which all the risk factors were recorded.

## 4.2 Choice of Transformation of Risk Factors

The risk factors considered in this analysis are the classical risk factors for BC, alcohol consumption and fat intake (total fat or saturated fat). There is strong prior evidence for a link between BC and classical risk factors such as age, menopausal status, age at menarche, parity, family history of BC, history of benign breast disease, educational level characterised by age at the end of schooling and Quetelet's body mass index (weight in Kg / height$^2$ in meters). It is thus necessary to control for the effect of these risk factors in any analyses of the role of alcohol and fat intake. This will be done by including classical risk factors in all the subsequent models. Nevertheless there is no consensus on whether some of these risk factors should be treated as continuous or categorical variables and, if categorical, on the most relevant class limits. A similar question arises for alcohol and fat intake.

As outlined in Section 3.2, we investigated suitable transformations of all these variables by an ACE analysis with monotonicity constraints. The variables were included in the ACE fit without any categorization and simultaneously (with only one of total fat or saturated fat).

The plots of the transformed variables against the original ones are shown in Figure 1. The only risk factor which clearly exhibits a linear pattern is age, and so age will not be transformed and will be entered in all the models as a continuous variable. For parity, a dichotomy is apparent and parity is transformed into a categorical variable : 3 or less children versus 4 or more children.

For age at menarche, age at the end of schooling, total and saturated fat intakes, simple linear transformations with thresholds seem to be suggested ; that is the transformed variable $X^*$ is equal to the original variable $X$ when $X \leq a$ while $X^*$ is equal to $a$ when $X > a$. For age at menarche, $a$=11.5 years was chosen ; for age at the end of schooling, $a$=16 years ; for total fat, $a$=800 grams per week ; for saturated fat, $a$=250 grams per week. For Quetelet's index, 2 thresholds are indicated with $X^*=X$ when $X \leq 20$, $X^*=20$ when $20 \leq X \leq 25.5$, $X^*=25.5$ when $X > 25.5$. Finally an interesting phenomenon appears when looking at alcohol consumption. The relationship seems fairly linear among the drinkers but there is a noticeable jump between the initial point representing the group of non drinkers (about half the sample) and the low drinkers. Hence, apart from treating alcohol consumption linearly, a dichotomous variable distinguishing drinkers from non drinkers will also be considered.

Some of the thresholds found above have an epidemiological interpretation. Early menarche is usually considered a detrimental risk factor whereas high parity would be protective. Here it is found that the risk seems to vary exclusively among very early menarche. The plot

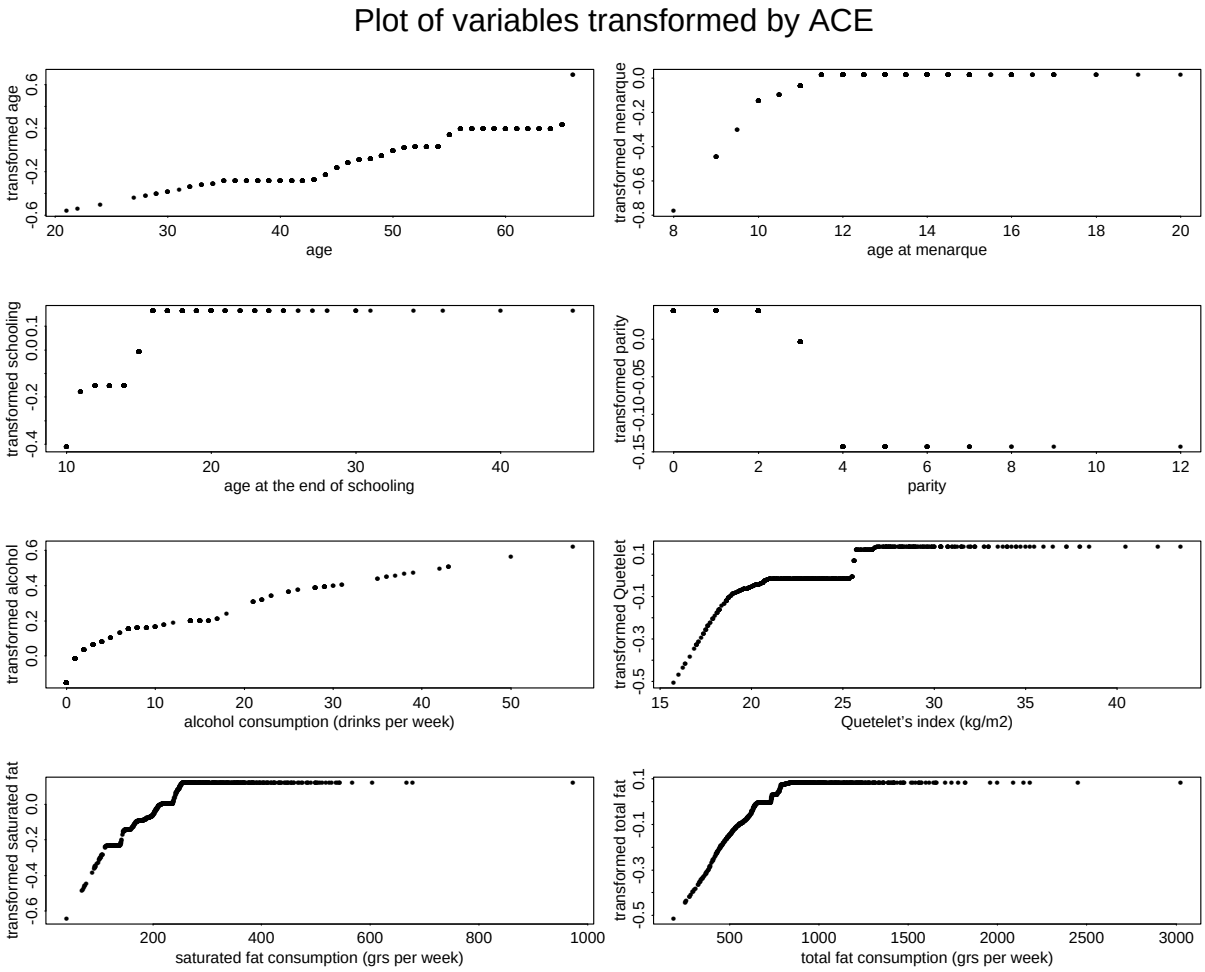


Figure 1: ACE transformations of the risk factors.

for parity highlights a break when women have 4 or more children, a fact that had been noticed in previous analyses of these data. Similarly, BC risk is known to increase with higher socio-economic status, and women prolonging their secondary education after age 16 (which is a natural stopping point in the French educational system) would tend to have different careers than those less well educated. The two thresholds found for Quetelet's index identify a group of slim women ($Q \leq 20$) and a group of corpulent women ($Q \geq 25$). For the nutritional variables, it is important to keep in mind that the thresholds do not necessary relate to an underlying biological phenomenon but could be connected to measurement error. For example the distinction between drinkers and non drinkers is less subject to measurement error than the quantitative measurement of alcohol consumption.

## 4.3 Choice of Independent Variables

From now on, in line with our prior knowledge, all the models considered include the classical risk factors transformed as suggested by the ACE plots (see Section 2.6).

Our first step is to study the effect of alcohol and fat intake, singly or jointly, on the risk of BC using the transformed variables. Alcohol intake is thus characterised by 2 variables : "alc" which is equal to the number of alcoholic drinks per week (the alcohol content - calculated using volume and average percent pure alcohol- is nearly identical for each type of drink) and "alc0" which dichotomises drinkers and non drinkers. Fat intake is represented by "tfat" and "tsat", the transformed threshold variables measuring consumption of these nutrients per week (in grams).

The dichotomous inclusion of each of these 4 variables leads to 16 models which were compared using GLIB (Table 1). Let us first look at the column $\phi = 1.65$ which is the recommended value for the prior variance parameter of the regression parameters (see Section 2.5). The highest value for $2\log B_{10}$ is obtained for model $M_{10}$. This value will be used as a "yardstick" and our first selection rule is to exclude all models which predicts the data far less well than the best model $M_{10}$. Applying the first rule of the Occam's Window strategy of Section 3.1, we excluded $M_k$ for which $\mathrm{pr}(M_k|D)/\mathrm{pr}(M_{10}|D) \leq 0.05$, i.e. for which the difference between $2\log B_{10}$ for the best model $M_{10}$ and $2\log B_{10}$ for model $M_k$ was greater than 6. This excluded models #1,#2,#4,#5,#11,#14,#15,#16.

Our second rule of model selection is to compare nested models and to exclude a model involving more parameters if a simpler model has a higher Bayes factor. We note that model #9 is nested in model #12 and has a higher Bayes factor, so model #12 is excluded. By a similar reasoning model #13 is excluded by reference to model #10 and #6 by reference to

Table 1: Selection of independent variables. All models include the classical risk factors (age, menopausal status, age at menarche, familial BC, history of benign breast disease).

| Model | alc | alc0 | tfat | tsat | $2 \log B_{10}$ | | | Deviance | df |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | $\phi$=1.00 | $\phi$=1.65 | $\phi$=5.00 | | |
| #1 | 0 | 0 | 0 | 0 | 0.0 | 0.0 | 0.0 | 1107.3 | 844 |
| #2 | 1 | 0 | 0 | 0 | 14.6 | 13.6 | 11.4 | 1086.2 | 843 |
| #3 | 0 | 1 | 0 | 0 | 18.9 | 17.9 | 15.7 | 1081.8 | 843 |
| #4 | 0 | 0 | 1 | 0 | 6.1 | 5.1 | 2.9 | 1094.6 | 843 |
| #5 | 0 | 0 | 0 | 1 | 5.9 | 4.9 | 2.7 | 1094.7 | 843 |
| #6 | 1 | 1 | 0 | 0 | 16.8 | 14.8 | 10.3 | 1077.6 | 842 |
| #7 | 1 | 0 | 1 | 0 | 17.3 | 15.3 | 10.9 | 1076.8 | 842 |
| #8 | 1 | 0 | 0 | 1 | 18.3 | 16.4 | 11.9 | 1075.7 | 842 |
| #9 | 0 | 1 | 1 | 0 | 21.7 | 19.7 | 15.3 | 1072.4 | 842 |
| #10 | 0 | 1 | 0 | 1 | 22.5 | 20.5 | 16.1 | 1071.6 | 842 |
| #11 | 0 | 0 | 1 | 1 | 1.4 | -0.6 | -5.0 | 1093.7 | 842 |
| #12 | 1 | 1 | 1 | 0 | 18.7 | 15.7 | 9.1 | 1069.0 | 841 |
| #13 | 1 | 1 | 0 | 1 | 19.8 | 16.9 | 10.2 | 1067.8 | 841 |
| #14 | 1 | 0 | 1 | 1 | 13.2 | 10.2 | 3.5 | 1075.5 | 841 |
| #15 | 0 | 1 | 1 | 1 | 17.4 | 14.4 | 7.8 | 1071.3 | 841 |
| #16 | 1 | 1 | 1 | 1 | 14.6 | 10.6 | 1.8 | 1067.6 | 840 |

Table 2: Model posterior probabilities for the 5 models selected by Occam's Window. All models include the classical risk factors (see Table 1).

| Model | Independent variable | Posterior probabilities (%) | | |
|---|---|---|---|---|
| | | $\phi$=1.00 | $\phi$=1.65 | $\phi$=5.00 |
| #3 | alc0 | 8 | 13 | 31 |
| #7 | alc, tfat | 4 | 4 | 3 |
| #8 | alc, tsat | 6 | 6 | 5 |
| #9 | alc0, tfat | 33 | 31 | 25 |
| #10 | alc0, tsat | 49 | 46 | 37 |

#3. We are thus left with 5 models : #3,#7,#8,#9,#10.

This model selection is robust to changes in the value of $\phi$, the parameter characterising the prior variance of the regression parameters. The same 5 models would remain whether we apply our selection rule for $\phi$ =1 or $\phi$ = 5. The values of the Bayes factor are indeed sensitive to the choice of $\phi$ but the relative rankings of the models are comparable across $\phi$. There is some evidence of a small amount of overdispersion, with an estimated overdispersion parameter of about $\tilde{\sigma} = 1.13$. Taking account of this would make relatively little difference to the results and we have not done so. However, it can be done using our approach; see Section 2.4.

Note that classical model comparisons between nested models using deviance differences would lead to somewhat different conclusions. For example, if we compare models #3 and #10 which are nested, the difference in deviance is 10.2 for 1 df which classically would give an overwhelming preference for model #10, that is to say that saturated fat has an effect, above that of alcohol. On the other hand the posterior probabilities which can be calculated from the Bayes factor indicate that model #10 is about 3.5 times more probable than model #3. So from a Bayesian point of view, the evidence for an effect of saturated fat, while positive, is not overwhelming.

Our next step was to rerun GLIB, including only the 5 previously selected models, in order to assess clearly their respective posterior probabilities (Table 2). Note that these posterior probabilities allow us to compare non nested models. Overall we see that the models including alc0 have higher posterior probabilities. An influence of the chosen values for $\phi$ is also apparent, with higher values of $\phi$ favouring smaller models, particularly model #3 which has fewer parameters. This is as expected; see Section 2.5.

The posterior probabilities that the regression coefficient is different from zero for each of

Table 3: Posterior probabilities for the inclusion of each independent variable.

| Independent variable | Posterior probabilities (%) | | |
|---|---|---|---|
| | $\phi$=1.00 | $\phi$=1.65 | $\phi$=5.00 |
| alc | 10 | 9 | 7 |
| alc0 | 90 | 91 | 93 |
| tfat | 37 | 35 | 27 |
| tsat | 55 | 52 | 41 |

the 4 variables included in the 5 models, which is also an output of GLIB, are given in Table 3. These probabilities are not much influenced by the values of $\phi$. All 5 retained models include either alc0 or alc as a variable, so that the evidence for an association between BC and alcohol consumption is very strong. The evidence for the part played by fat consumption (either total fat or saturated fat only) is less strong. There are odds of only 7:1 ($\phi$=1.65) supporting it. We can also compare the qualitative and quantitative effects of alcohol: there are odds of 9:1 in favour of the main effect of alcohol being of the dichotomous type. On the other hand, there is only weak evidence for the effect of fat consumption to be restricted to saturated fat since the odds are only roughly 1.5:1 in favour of saturated fat. This is an area where substantial uncertainty remains.

## 4.4 Confirmatory Analysis of the Choice of Transformations

The appropriateness of the choice of transformations for total fat and saturated fat can be confirmed by comparing the chosen models to similar models where the variables have their original quantitative values. This was done systematically by including variables without thresholds in models #7 to #10 (Table 4). Furthermore, a variable alc3 discretising alcohol consumption into 3 classes (0, 1-7, $>$ 7) was also created and compared to the dichotomised alc0 variable. The results in Table 4 indicate without exception that the models including the transformed variables with thresholds have substantially higher posterior probabilities. The variable alc3 does not perform well either.

Thus the coding chosen on the basis of ACE as in Section 4.2 is adequate. If it had not been, we would have iterated as in Section 3.3, changing the codings and redoing the independent variable selection.

Table 4: Confirmatory analysis of the choice of transformations of independent variables via Bayes factors ($\phi = 1.65$).

| Model | Independent variables | $2\log B_{10}$ |
|---|---|---|
| #3 | alc0 | 17.9 |
| #7 | alc, tfat | 15.3 |
| | alc, fat | 8.4 |
| #8 | alc, tsat | 16.4 |
| | alc, sat | 11.1 |
| #9 | alc0, tfat | 19.7 |
| | alc0, fat | 12.4 |
| #10 | alc0, tsat | 20.5 |
| | alc0, sat | 14.8 |
| | alc3, tsat | 14.4 |

Table 5: Posterior means of the regression coefficients under the 5 models ($\phi = 1.65$).

| Model | Variable | | | |
|---|---|---|---|---|
| | alc0 | alc | tfat | tsat |
| #3 | .734 | – | – | – |
| #7 | – | .038 | .0017 | – |
| #8 | – | .039 | – | .0052 |
| #9 | .691 | – | .0017 | – |
| #10 | .704 | – | – | .0051 |

## 4.5 Model Uncertainty

In our example the regression coefficients estimated under the different models do not vary much (Table 5). Furthermore they are similar to the maximum likelihood estimators (results not shown). For example the coefficient for the variable alc0 given by GLIB ($\phi = 1.65$) varies between 0.73 (model #3), 0.69 (model #9) and 0.70 (model #10) and is equal to the corresponding maximum likelihood estimate up to two decimal places. This lack of variation of coefficients accross models is not a general occurrence. When some but not all the models considered include interaction terms, the regression coefficients estimated can be substantially different, as was the case in the study of oral contraceptive use and myorcardial infarction analyzed in Raftery (1993).

In this study, the main consequences of model uncertainty are thus concentrated on

evaluating the risk for particular individuals. For example, for a women having 3 alcoholic drinks per week, models #7 and #8 estimate odds ratios values around 1.1 whereas models #3, #9 and #10 lead to odds ratios values around 2. Overall, an odds ratio of 1.9 for 3 drinks per week would be estimated if the odds ratio for each model are weighted by the corresponding model probabilities. On the other hand, for a women having 18 drinks per week, the odds ratio for alcohol consumption would be equal under all the models.

There is a strong correlation between overall fat intake and saturated fat intake ($r = 0.89$), with the intakes related usually by a ratio of 3.2. This is well reflected by the ratio of the corresponding regression coefficients. So for women whose consumption of saturated fat is about a third of their overall fat consumption, odds ratio estimates given by models #7 to #10 will be comparable. On the other hand, there are a few women consuming a much higher proportion of saturated fat, for example patient 41 consumed 390 grs per week of saturated fat and 645 grs per week of total fat. For these women, the odds ratio given by model #9 is 3 while that given by model #10 is 3.6. Note that this discrepancy would have considerably widened if the threshold of 250 had not been used for saturated fat.

## 4.6 Comments

In the original analysis of this study, an elevation of BC risk associated with alcohol consumption, categorized either in 5 classes (0, 1-2, 3-9,10-17,$>$ 17) or 3 classes (0, 1-7, $>$ 7), had been found. The effect of fat consumption was not apparent unless a negative interaction (significant at the 5% level) was introduced in the model. It had also been noted that the odds ratio (1.8) corresponding to the lowest class of alcohol consumption (1-2 drinks per week) was higher than would be expected if the risk increased linearly with consumption of alcohol.

In the new analysis that we present here, thresholds were strongly indicated for the effect of fat consumption. With the introduction of these thresholds, there is substantial but not overwhelming evidence for an effect of fat consumption, whether total or restricted to saturated fat. No interaction comes into play. This was checked by calculating Bayes factor for models including interactions between alcohol and fat consumption. In essence, the negative interactions, which were somewhat unnaturally introduced in the previous analyses, are superceded by the use of variables with thresholds.

The new analysis has also given thought-provoking results about alcohol consumption. Indeed, there is clear evidence that models including the simple dichotomy, drinker versus non-drinker, are preferred. From a biological point of view, this is a surprising result since

none of the several biological hypotheses which have been put forward to explain the role of ethanol in mammary carcinogenesis could account for such a threshold. Hence we would favour an alternative explanation of this result ; that the notorious difficulties in measuring alcohol consumption have resulted in imprecise measurements of the quantity of alcohol consumed to such a degree that the dichotomy, drinker versus non-drinker, better explains the variation in BC risk in our data.

In conclusion, by quantifying the model uncertainties well, areas where future research is needed are more clearly identified. There is a need to improve the assessment of alcohol consumption, and dietary history questionnaires could usefully be complemented by other measuring instruments. It was not our purpose to assess in detail the specific role which could be played by different components of fat intake. Let us just say that, as our results stand, substantial model uncertainty remains concerning a specific role of saturated fat intake and that populations where there is less correlation between total fat and saturated fat intakes should be sought after.

# 5 The GLIB Software

GLIB is an S-PLUS function for implementing the methods that we have described. It can be obtained free of charge from StatLib by sending the e-mail message "send glib from S" to *statlib@stat.cmu.edu*. Here are some of its features:

- For each model considered, GLIB returns the posterior model probability and the Bayes factor for it against the null model (i.e. the one with no independent variables).
- For each independent variable, GLIB returns the posterior probability that the corresponding regression parameter is nonzero (obtained by summing across models), and its composite posterior mean and standard deviation given that it is nonzero, from equations (5) and (6).
- GLIB uses the reference proper priors of Section 2.5 and the user can specify a set of values of the prior dispersion parameter $\phi$. The Bayesian results are returned for each value of $\phi$, thus assessing sensitivity to the prior. The default values are $\phi$=1.00, 1.65, 5.00, i.e. the lower bound, the recommended value and the upper bound from Section 2.5. The numerical values of the prior mean and variance matrix used for each model and each value of $\phi$ are output on request by specifying `priorvar=T`.

- The user can specify the prior model probabilities via the input `pmw`. Actually, these are prior model *weights* and do not have to sum to one; GLIB renormalizes them automatically. This is convenient if the program is being rerun after removing some of the models initially considered. By default, all models are assumed equally likely *a priori*.

- GLIB returns standard frequentist GLIM results (deviances, MLEs, standard errors, and so on), as well as their Bayesian analogues. This allows an arbitrary number of generalized linear models to be estimated with a single command, which can be useful even if the Bayesian aspects are not of primary interest.

The models to be fit are specified by the input design matrix `x` and the input matrix `models`, which specifies the subsets to be considered. Columns 2–5 of Table 1 constitute an example of the input `models` matrix.

# 6 Discussion

We have described a Bayesian model-building strategy for generalized linear models that avoids the difficulties of commonly-used *ad-hoc* strategies, namely arbitrariness in the coding of risk factors, lack of knowledge of the overall properties of model selection strategies based on significance tests, and failure to account for model uncertainty. It delivers a parsimonious set of models together with their posterior probabilities, thereby facilitating effective communication of model uncertainty. As a result, the conclusions from the analysis are clear, as are the remaining uncertainties. This helps to define future research priorities.

In our strategy, model selection is based on Bayes factors rather than on significance tests. This has been argued to be a more appropriate form of inference (Edwards, Lindeman and Savage, 1963; Berger and Delampady, 1987), and it tends to give more reasonable results in large samples (e.g. Raftery, 1986), as well as allowing easy comparison of non-nested models (Kass and Raftery, 1994). The methods can be implemented using the GLIB software available free of charge from StatLib. This software automatically assesses the sensitivity of the results to the prior dispersion.

The strategy yields the posterior probability of inclusion for each risk factor considered, and thus answers what is often the main question asked in epidemiological studies, namely, how sure can we be that there is an association between risk factor and disease? It also provides a formal justification for including "classical" risk factors even when the data at

hand provide little support for their inclusion. The justification is that the evidence for their inclusion comes from prior studies and is reflected in low prior model probabilities for models that exclude them.

These points are well illustrated by the epidemiological application of Section 4. Previous analyses using a more standard strategy had led to somewhat ambiguous and counter-intuitive conclusions, particularly about whether breast cancer is associated with fat consumption (see Section 4.6). Our analysis, by contrast, revealed positive (but not decisive) evidence for such an association. It is clear also that the data do not allow us to determine whether the effect is due to total fat or just to saturated fat. Our analysis also showed clearly that alcohol consumption is a risk factor for breast cancer, although the amount of alcohol consumed is poorly measured. Thus future research that replicates this study should focus on better measurement of alcohol consumption and on finding populations where it is possible to distinguish between the effects of different types of fat consumption.

Accurate Bayesian analysis is possible for generalized linear models because a good approximation is available for the Bayes factors via the Laplace method. That may not be the case for other biostatistical models, although approximations of similar quality are available for some hierarchical models used in health services research: beta-binomial models (Kahn and Raftery, 1991) and Poisson-Gamma models (Rosenkranz, 1992). For other biostatistical models for which such approximations are not yet available, it is often possible to use the BIC or Schwarz approximation (Schwarz, 1978), namely

$$\log \mathrm{pr}(D|M_k) \sim \log \mathrm{pr}(D|\hat{\theta}_k, M_k) - \tfrac{1}{2} p_k \log n.$$

On the face of it, this is a crude approximation, but it has been observed to be surprisingly accurate. Kass and Wasserman (1992) have pointed out that it is, in fact, quite an accurate approximation for a particular, reasonable, prior. This opens up a wide class of models for the kind of model selection strategy and model uncertainty analysis outlined in this chapter.

# References


Berger, J.O. and Delampady, M. (1987). Testing precise hypotheses (with Discussion). *Statistical Science*, 3, 317-352.

Berger, J.O. and Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of $P$ values and evidence (with Discussion). *Journal of the American Statistical Association,* 82, 112-122.

Breiman, L. and Friedman, J.H. (1985). Estimating optimal transformations for multiple regression and correlation (with Discussion). *Journal of the American Statistical Association*, 80, 580–619.

Breslow, N.E. and Day, N.E. (1980). *Statistical Methods in Cancer Research. Volume I — The Analysis of Case-Control Studies.* Lyon, France: International Agency for Research on Cancer.

de Bruijn, N.G. (1970). *Asymptotic Methods in Analysis.* Amsterdam: North-Holland.

DeVeaux, R.D. (1989). Finding transformations for regression using the ACE algorithm. *Sociological Methods and Research*, 18, 327–359.

Edwards, W., Lindman, H. and Savage, L.J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70, 193–242.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models.* London: Chapman and Hall.

Howe, G., Rohan, T. and Decarli, A. *et al.* (1991). The association between alcohol and breast cancer risk : evidence from the combined analysis of six dietary case-control studies. *International Journal of Cancer*, 47, 707–710.

Kahn, M.J. and Raftery, A.E. (1991). Discharge rates of Medicare stroke patients to skilled nursing facilities: Bayesian logistic regression with unobserved heterogeneity. Technical Report, Department of Statistics, University of Washington.

Kass, R.E. and Raftery, A.E. (1994). Bayes factors. *Journal of the American Statistical Association*, to appear.

Kass, R.E. and Wasserman, L. (1992). The surprising accuracy of the Schwarz criterion as an approximation to the log Bayes factor. Technical Report, Department of Statistics, Carnegie-Mellon University.

Kass, R.E. and Vaidyanathan, S. (1992). Approximate Bayes factors and orthogonal parameters, with application to testing equality of two Binomial proportions. *Journal of the Royal Statistical Society, series B*, 54: 129-144.

Madigan, D. and Raftery, A.E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, to appear.

Madigan, D. and York, J. (1993) Bayesian graphical models for discrete data. Technical Report no. 259, Department of Statistics, University of Washington.

Madigan, D., Raftery, A.E., York, J.C., Bradshaw, J.M. and Almond, R.G. (1993). Strategies for graphical model selection. In *Proceedings of the 4th International Workshop on Artificial*

*Intelligence and Statistics*, to appear.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models* (2nd ed.). London: Chapman and Hall.

Racine, A., Grieve, A.P., Fluhler, H., and Smith, A.F.M. (1986). Bayesian methods in practice: experiences in the pharmaceutical industry (with Discussion). *Applied Statistics,* 35, 93-150.

Raftery, A.E. (1986). Choosing models for cross-classifications. *American Sociological Review*, 51, 145–146.

Raftery, A.E. (1988). Approximate Bayes factors for generalized linear models. Technical Report no. 121, Department of Statistics, University of Washington.

Raftery, A.E. (1992). Bayesian model selection in structural equation models. In *Testing Structural Equation Models* (eds. K.A. Bollen and J.S. Long), Beverly Hills: Sage.

Raftery, A.E. (1993). Approximate Bayes factors and accounting for model uncertainty in generalized linear models. Technical Report no. 255, Department of Statistics, University of Washington.

Raftery, A.E., Lewis, S.M. and Aghajanian, A. (1993). Demand or ideation? Evidence from the Iranian marital fertility decline. Working Paper, Center for Studies in Demography and Ecology, University of Washington.

Raftery, A.E., Lewis, S.M., Aghajanian, A. and Kahn, M.J. (1992). Event history modeling of World Fertility Survey data. Working Paper no. 93-1, Center for Studies in Demography and Ecology, University of Washington.

Raftery, A.E., Madigan, D.M. and Hoeting, J. (1993). Model selection and accounting for model uncertainty in linear regression models. Technical Report no. 262, Department of Statistics, University of Washington.

Richardson, S., de Vincenzi, I., Gerber, M. and Pujol, H. (1989). Alcohol consumption in a case-control study of breast cancer in Southern France. *International Journal of Cancer*, 44, 84–89.

Richardson, S., Gerber, M., and Cénée S. (1991). The role of fat, animal protein and some vitamin consumption in breast cancer : a case-control study in Southern France. *International Journal of Cancer*, 48, 1–9.

Rosenkranz, S. (1992). The Bayes factor for model evaluation in a hierarchical Poisson model for area counts. Ph.D. dissertation, Department of Biostatistics, University of Washington, 1992.

Ségala, C., Gerber, M. and Richardson S. (1991). The pattern of risk factors for breast cancer in a Southern France population: Interest for a stratified analysis by age at diagnosis. *British*

*Journal of Cancer*, 64, 919–925.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics,* 6, 461-464.

Tierney, L., and Kadane, J.B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81, 82-86.

Van der Brandt, P.A., Van't Veer, P., Goldbohm, R.A. *et al.* (1993). A prospective cohort study on dietary fat and the risk of postmenopausal breast cancer. *Cancer Research*, 53, 75–82.