Chapter 13

MODEL AVERAGING

by Merlise Clyde[1]

13.1 INTRODUCTION

In Chapter 12, we considered inference in a normal linear regression model with q predictors. In many instances, the set of predictor variables $\mathbf{X}$ can be quite large, as one considers many potential variables and possibly transformations and interactions of these variables that may be relevant to modelling the response $\mathbf{Y}$. One may start with a large set to reduce chances that an important predictor has been omitted, but employ variable selection to eliminate variables do not appear to be necessary and avoid over-fitting. Historically, variable selection methods, such as forwards, backwards, and stepwise selection, maximum adjusted R^2, AIC, Cp, etc., have been used, and, as is well known, these can each lead to selection of a different final model (Weisberg, 1985). Other modeling decisions that may arise in practice include specifying the structural form of the model, including choice of transformation of the response, error distribution, or choice of functional form that relates the mean to the predictors. Decisions on how to handle "outliers" may involve multiple tests with a somewhat arbitrary cut-off for p-values or the use of "robust" outlier resistant methods. Many of the modeling decisions are made conditional on previous choices, and final measures of "significance" may be questionable.

While one may not be surprised that approaches for selection of a model reach different conclusions, a major problem with such analyses is that often only a "best" model and its associated summaries are presented, giving a false impression that this is the only model that explains the data. This

[1]Institute of Statistics and Decision Sciences, Duke University, Durham, NC. Email: clyde@stat.duke.edu

standard practice ignores uncertainty due to model choice, and can lead to over-confident inferences and predictions, and decisions that are riskier than one believes (Draper, 1995; Hodges, 1987).

In this chapter, we review a Bayesian approach to address model uncertainty known as Bayesian Model Averaging, and describe its use in linear regression models for the choice of covariates and generalized linear models for the choice of link function and covariates.

13.2 MODEL AVERAGING AND SUBSET SELECTION IN LINEAR REGRESSION

Let $\mathbf{X}$ denote the n by q matrix of all predictors under consideration. Under the full model (all predictors), the univariate multiple regression model is represented as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

where $\mathbf{u} \sim N(0, \sigma^2 I)$. In the problem of subset or variable selection among the q predictor variables, models under consideration correspond to potentially all possible subsets of the q variables, leading to a model space $\mathcal{M} = \{\mathcal{M}_1, \ldots, \mathcal{M}_K\}$ where $K = 2^q$ and includes the model with no predictor variables at all. Models for different subsets may be represented by a vector of binary variables, $\boldsymbol{\gamma} = (\gamma_1, \ldots \gamma_q)'$ where γ_j is an indicator for inclusion of variable $\mathbf{X}_j$ under model $\mathcal{M}_k$. A convenient indexing of models in the all subset regression problem is to let $\boldsymbol{\gamma}$ denote the binary representation of k for model $\mathcal{M}_k$. Under $\mathcal{M}_k$ there are $q_{\boldsymbol{\gamma}} = \sum_{j=1}^{q} \gamma_j$ non-zero parameters, $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$, with $q_{\boldsymbol{\gamma}} \times n$ design matrix $\mathbf{X}_{\boldsymbol{\gamma}}$.

To incorporate model uncertainty regarding the choice of variables in the linear regression model, we build a hierarchical model (George and McCulloch, 1993, 1997; Raftery et al., 1997):

$$\begin{aligned}
\mathbf{Y}|\boldsymbol{\beta}, \sigma^2, \mathcal{M}_k &\sim N(\mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2 I_n) && (13.1)\\
\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\sigma^2, \mathcal{M}_k &\sim p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\mathcal{M}_k, \sigma^2) && (13.2)\\
\sigma^2|\mathcal{M}_k &\sim p(\sigma^2|\mathcal{M}_k) && (13.3)\\
\mathcal{M}_k &\sim p(\mathcal{M}_k) && (13.4)
\end{aligned}$$

where the first stage is based on the normal probability model of Chapter 12 using $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ and the design matrix $\mathbf{X}_{\boldsymbol{\gamma}}$ under model $\mathcal{M}_k$. Variables are

excluded in model $\mathcal{M}_k$ by setting elements of the full parameter vector $\boldsymbol{\beta}$ to zero. The second stage corresponds to a prior distribution for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$, the non-zero elements of $\boldsymbol{\beta}$, under each model. The resulting distribution for the q dimensional vector of regression parameters $\boldsymbol{\beta}$ can be viewed as a mixture of a point masses at zero and continuous distributions for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$. The last stage of the hierarchical model assigns prior probabilities over all models under consideration, conditional on the model space $\mathcal{M}$.

The posterior distribution over models in $\mathcal{M}$ is

$$p(\mathcal{M}_k|\mathbf{Y}) = \frac{m(\mathbf{Y}|\mathcal{M}_k)p(\mathcal{M}_k)}{\sum_k m(\mathbf{Y}|\mathcal{M}_k)p(\mathcal{M}_k)}$$

where $m(\mathbf{Y}|\mathcal{M}_k)$ is the marginal distribution of the data under model $\mathcal{M}_k$,

$$m(\mathbf{Y}|\mathcal{M}_k) = \int\int p(\mathbf{Y}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2, \mathcal{M}_k)p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\sigma^2, \mathcal{M}_k)p(\sigma^2|\mathcal{M}_k)d\boldsymbol{\beta}_{\boldsymbol{\gamma}}d\sigma^2$$

obtained by integrating over the prior distributions for model specific parameters. The posterior distribution $p(\mathcal{M}_k|\mathbf{Y})$ provides a summary of model uncertainty after observing the data $\mathbf{Y}$. For the more general problem of model selection, the marginals are obtained similarly by integrating over all model specific parameters.

If Δ is a quantity of interest, say predicted values at a point x, then the expected value of Δ given the data $\mathbf{Y}$ is obtained by first finding the posterior expectation of Δ under each model, and then weighting each expectation by the posterior probability of the model:

$$E(\Delta|\mathbf{Y}) = \sum_k p(\mathcal{M}_k|\mathbf{Y})E(\Delta|\mathcal{M}_k, \mathbf{Y}). \tag{13.5}$$

Similarly, the posterior distribution for Δ can be represented as a mixture distribution over all models,

$$p(\Delta|\mathbf{Y}) = \sum_k p(\mathcal{M}_k|\mathbf{Y})p(\Delta|\mathbf{Y}, \mathcal{M}_k) \tag{13.6}$$

where $p(\Delta|\mathbf{Y}, \mathcal{M}_k)$ is the posterior distribution of Δ under model $\mathcal{M}_k$. As an alternative to p-values, a useful summary is the marginal posterior probability that $\beta_j = 0$, which is equivalent to $P(\gamma_j = 0|\mathbf{Y})$. This can be obtained by summing the posterior model probabilities over all models where γ_j is zero.

Likewise, the posterior probability that β_j is not zero (or that variable $\mathbf{X}_j$ has an effect on $\mathbf{Y}$) can be obtained as $1 - P(\gamma_j = 0|\mathbf{Y})$.

The Bayesian solution for incorporating model uncertainty has become known as Bayesian Model Averaging (BMA) (Hoeting et al., 1999) as quantities of interest can often be expressed as a weighted average of model specific quantities, where the weights depend on how much the data support each model (as measured by the posterior probabilities on models). If the posterior probability is concentrated on a single model, then model uncertainty is not an issue and both model selection and model averaging will lead to similar results. In many cases, model uncertainty dominates other forms of uncertainty, such as parameter uncertainty and sampling variation, and BMA can lead to real improvements in predictive performance (see Hoeting et al. (1999) for several examples illustrating BMA).

On the surface, Bayesian model averaging and model selection are straightforward to implement: one specifies the distribution of the data under each model, and the prior probabilities of models and model specific parameters; Bayes theorem provides the rest. The two major challenges confronting the practical implementation of Bayesian model averaging are choosing prior distributions and calculating posterior distributions. The latter problem requires that one can actually carry out the integration necessary for obtaining the marginal distribution of the data $m(\mathbf{Y}|\mathcal{M}_k)$, as well as determining the normalizing constant in the denominator of the posterior probability if M_k. In the normal linear regression model, there are many choices of prior distributions that lead to closed form solutions for the marginal distribution (George and McCulloch, 1997). If the number of models in $\mathcal{M}$ is too large to permit enumeration of all models, one can approximate BMA using a subset of models. Where integrals can not be carried out analytically, there are a range of methods that can be used to implement BMA, from asymptotic approximations to reversible jump Markov chain Monte Carlo sampling (see review articles by Hoeting et al. (1999); Chipman et al. (2001) for an overview).

13.3 PRIOR DISTRIBUTIONS

The specification of prior distributions is often broken down into two parts: (1) elicitation of distributions for parameters specific to each model, such

as the distribution for regression coefficients in linear models, $p(\boldsymbol{\beta}_k|\mathcal{M}_k, \sigma^2)$, and (2) selection of a prior distribution over models $p(\mathcal{M}_k)$. For moderate to high dimensional problems, it is difficult to consistently specify separate prior distributions for parameters under each model and elicit probabilities of each $\mathcal{M}_k$ directly. Practical implementations of BMA have usually made prior assumptions to simplify prior elicitation and allow tractable computations using conjugate prior distributions.

13.3.1 Prior Distributions on Models

In the context of variable selection, one may specify a prior distribution over $\mathcal{M}$ by a probability distribution on the indicator variables $\boldsymbol{\gamma}$. In most applications to date, the indicator variables are taken to be independent a priori, using a product of independent Bernoulli distributions,

$$P(\boldsymbol{\gamma}_k) = \prod_q \omega_j^{\gamma_{jk}} (1 - \omega_j)^{1-\gamma_{jk}}$$

for the prior distribution of $\boldsymbol{\gamma}$. The hyperparameter ω_j corresponds to the prior probability that variable $\mathbf{X}_j$ is included. The choice of ω_j could be based on subjective information, or could be assigned a prior distribution, such as a beta distribution, reflecting additional uncertainty. As a special case of the independent prior distribution over models, the uniform distribution over models ($\omega_j = 0.5$) is often recommended as a default choice. The uniform distribution is appealing in that posterior probabilities of models depend only on the marginal likelihood of the data, and not prior probabilities of models. One should note, however, that the uniform distribution implies that the model size has a prior distribution that is Binomial$(q, 0.5)$ and a prior belief that half of the variables are expected to be included, which may not be realistic in problems where q is large. For models that contain interactions, polynomial terms or highly correlated variables, independent prior distributions may not be suitable, and a prior that takes into account the dependence structure may be preferred (Chipman, 1996). Prior distributions over models that account for expected correlations between variables is an area of ongoing research which should lead to more realistic prior distributions over models (see Chipman et al. (2001) for new directions).

13.3.2 Prior Distributions for Model Specific Parameters

In the context of the linear regression model, by far the most common choice for prior distributions on parameters within models, $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$, is a conjugate normal prior distribution

$$\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma} \sim N(0, \sigma^2 \boldsymbol{\Sigma}_{\boldsymbol{\gamma}})$$

(Chipman et al., 2001). As one cannot typically specify a separate prior distribution for $\boldsymbol{\beta}$ under each model, any practical implementation for Bayesian model averaging usually resorts to structured families of prior distributions. This also ensures that prior specifications for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ are "compatible" across models (Dawid and Lauritzen, 2001). To avoid incompatibilities between nested models, choices for $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}$ are obtained from the prior distribution for $\boldsymbol{\beta}$ under the full model and finding the conditional distribution for $\boldsymbol{\beta}$ given that a subset of $\boldsymbol{\beta} = 0$. For example, Zellner's g-prior (Zellner, 1986) is commonly used, which leads to $\boldsymbol{\Sigma} = g(\mathbf{X}'\mathbf{X})^{-1}$ for the full model and $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}} = g(\mathbf{X}'_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}})^{-1}$ for the coefficients in model $\boldsymbol{\gamma}$. This reduces the choice of hyperparameters down to just g, a scale parameter in the prior variance.

In order to obtain closed form marginal distributions for the data, a conjugate inverse gamma prior distribution for σ^2 is commonly used. This is equivalent to taking

$$\frac{\nu\lambda}{\sigma^2} \sim \chi^2_\nu \tag{13.7}$$

where ν and λ are fixed hyperparameters.

13.4 POSTERIOR DISTRIBUTIONS

Using a conjugate prior distributions for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ combined with the inverse gamma prior for σ^2, the marginal likelihood for $\mathbf{Y}$ is

$$m(\mathbf{Y}|\boldsymbol{\gamma}) \propto |\mathbf{X}'_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}} + \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1}|^{-1/2}|\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}|^{-1/2}(\lambda\nu + S^2_{\boldsymbol{\gamma}})^{-(n+\nu)/2} \tag{13.8}$$

where

$$S^2_{\boldsymbol{\gamma}} = \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}_{\boldsymbol{\gamma}}(\mathbf{X}'_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}} + \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1})^{-1}\mathbf{X}'_{\boldsymbol{\gamma}}\mathbf{Y}$$

is a Bayesian analog to the residual sum of squares. Under Zellner's g-prior the marginal simplifies further, leading to

$$m(\mathbf{Y}|\boldsymbol{\gamma}) \propto (1+g)^{-q_{\boldsymbol{\gamma}}/2}(\nu\lambda + \mathbf{Y}'\mathbf{Y} - \frac{g}{1+g}(\hat{\boldsymbol{\beta}}'_{\boldsymbol{\gamma}}(\mathbf{X}'_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}})\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}})^{-(n+\nu)/2} \tag{13.9}$$

where $\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ is the ordinary least squares estimate of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ under model $\boldsymbol{\gamma}$ and $q_{\boldsymbol{\gamma}}$ is the number of regression parameters (including the intercept) in model $\boldsymbol{\gamma}$. This can be normalized

$$p(\boldsymbol{\gamma}|\mathbf{Y}) = \frac{m(\mathbf{Y}|\boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma}' \in \mathcal{M}} m(\mathbf{Y}|\boldsymbol{\gamma}')} \tag{13.10}$$

to obtain the posterior model probabilities which can be easily calculated from summary statistics using most regression packages.

13.5 CHOICE OF HYPERPARAMETERS

In order to implement BMA in the linear regression problem, even under structured prior distributions, there are still choices that must be made regarding the hyperparameters g, ν and λ, in addition to the parameters in $p(\boldsymbol{\gamma})$. Where possible, subjective considerations should be used, however, one may often want to present an "objective" summary in addition to subjective analyses. As a "default" choice, both Smith and Kohn (1996) and Fernández et al. (2001) have recommended using a uniform prior distribution for $\log(\sigma^2)$, corresponding to the limiting case of the conjugate prior as $\nu \to 0$. This is invariant to scale changes in $\mathbf{Y}$ and while improper, leads to proper posterior distributions.

Vague choices for g in Zellner's g-prior should be avoided, since being too "non-informative" about $\boldsymbol{\beta}$ by taking g large can have the unintended consequence of favoring the null model *a posteriori* Kass and Raftery (1995), as g has an influential role in the posterior model probabilities, as inspection of (13.8) reveals. Default choices for g can be calibrated based on information criteria such as AIC (Akaike Information Criterion (Akaike, 1973), see also Chapter 15), BIC (Bayes Information Criterion – (Schwarz, 1978), see also Chapter 15), or RIC (Risk Inflation Criterion –(Foster and George, 1994)). Based on simulation studies, Fernández et al. (2001) prefer RIC-like prior distributions when $n < p^2$ and BIC-like prior distributions otherwise, and recommend using $g = \max(n, p^2)$.

BIC can be rapidly calculated for a wide class of models, including generalized linear models where analytic integration to obtain the marginal likelihood

is not possible,

$$BIC(\mathcal{M}_k) = -2\log(\text{ maximized likelihood}|\mathcal{M}_k) + q_k \log(n) \qquad (13.11)$$

where q_k is the dimension of model $\mathcal{M}_k$. This can be used to obtain approximate posterior model probabilities where

$$p(\mathcal{M}_k|\mathbf{Y}) = \frac{p(\mathcal{M}_k)\exp(-.5BIC(\mathcal{M}_k))}{\sum_k p(\mathcal{M}_k)\exp(-.5BIC(\mathcal{M}_k))}.$$

Model probabilities based on BIC to permit a default objective analysis, have provided improved out of sample predictive performance in a wide variety of settings (Hoeting et al., 1999) and avoid the issue of hyperparameter specification. BIC may be conservative in problems with small to moderate sample sizes.

Empirical Bayes (EB) approaches provide an adaptive (but data based) choice for g (Clyde and George (2000); George and Foster (2000); Clyde (2001); Chipman et al. (2001), which typically have better out-of-sample performance than fixed hyperparameter specifications or using BIC, but with additional computational complexity. EB estimates of g can be obtained iteratively, and still permit analytic expressions for BMA without using Monte Carlo sampling in small to moderate problems where the model space can be enumerated or a subset of models can be identified using leaps and bounds (see below). This provides a compromise between fixed specifications for g and using a fully Bayesian approach with an additional stage in the hierarchical model for a prior distribution on g. While more computationally demanding, the use of a prior distribution on g, such as an inverse gamma distribution, provides additional prior robustness by inducing heavier tails in the prior distribution.

13.6 IMPLEMENTING BMA

In the variable selection problem for linear regression, marginal likelihoods are available in closed form (at least for nice conjugate prior distributions); for generalized linear models and many other models, Laplace's method of integration or BIC can provide accurate approximations to marginal distributions. The next major problem is that the model space is often too large to allow enumeration of all models, and beyond 20-25 variables, estimation

of posterior model probabilities, model selection, and BMA must be based on a sample of models.

Deterministic search for models using branch and bounds or leaps and bounds algorithms (Furnival and Wilson, 1974) is efficient for problems with typically fewer than 30 variables and is used in several programs such as `BICREG` and `BICGLM` (Hoeting et al., 1999). For larger problems, these methods are too expensive computationally or do not explore a large enough region of the model space. Gibbs and Metropolis Hastings MCMC algorithms (see Chapter 6) work well in larger problems for stochastically sampling models (see Hoeting et al. (1999); George and McCulloch (1993, 1997) for details on algorithms).

Given a subset of models obtained by deterministic or stochastic search, BMA is carried out by summing over the sampled models ($\mathcal{S}$) rather than the entire space $\mathcal{M}$ in equations (13.5), (13.6), (13.10) and (13.11).

13.7 EXAMPLES

We now look at two examples to illustrate model averaging. In both examples, the model spaces can be enumerated and probabilities were approximated by BIC using the output from linear model and generalized linear model software. Software for BMA using BIC and Gibbs sampling is available on the BMA website, URL: `www.research.att.com/~volinsky/bma.html`, maintained by Chris Volinsky.

13.7.1 Pollution and Mortality

There have been ongoing discussions about the effect of air pollution on mortality. Data in one early study (1960) was based on a cross-sectional sample of 60 metropolitan areas (data available in Ramsey and Schafer (2002)). The response variable is age-adjusted mortality from all causes (deaths per 100,000 population). Pollution indices include "relative pollution potential" for NOX (nitrogen dioxides), SO2 (sulfur dioxide) and HC (hydrocarbons), where the relative pollution potential is the product of tons emitted per day per square kilometer and a factor used to correct for the area and exposure in the metropolitan areas. Twelve potential confounding variables under consideration include PRECIP (mean annual precipitation), JANTEMP (mean

January temperature), JULYTEMP (mean annual temperature in July), OVER65 (percentage of population over age 65), HOUSE (population per household), EDUC (median number of years of school completed), SOUND (percentage of sound housing), DENSITY (population density), NONWHITE (percentage of population in 1960 that was non-white), WHITECOl (percentage of white collar occupations), and HUMIDITY (annual average relative humidity). Pollution concentrations were transformed using natural logarithms.

If uncertainty about whether all confounding variables and which of the pollution variables is related to mortality is taken into consideration, there are $2^{15} = 32,768$ potential models. All possible models were fit using ordinary least squares and approximate model probabilities based on BIC (13.11) were calculated for each model. Figure 13.1 illustrate the top 25 best models in terms of BIC. The columns in the image correspond to variables, with rows corresponding to models. White indicates variable exclusion, while dark rectangles indicate that the variable for that column is included for the model in that row. The y-axis corresponds to the intensity which is scaled to the log(Bayes Factor), the difference in BIC for comparing each model to the model with the largest BIC value. The best model in terms of BIC is at the top of the image. While log(NOX) is included in all of the top 25 models, there is more uncertainty about te importance of HC and SO2, and which of the other potential explanatory variables to include.

The log posterior odds that there is a pollution effect (one or more of the three indices is included) after adjusting for all potential confounders is 4.44. Using Jeffreys' scale of evidence (Kass and Raftery, 1995), applied to log posterior odds rather than Bayes Factors, this suggests positive evidence in favor of the hypothesis that pollution is associated with mortality. Posterior probabilities are preferable to Bayes factors, as Bayes factors are not necessarily monotonic for nested models (Lavine and Schervish, 1999).

Unlike p-values, marginal posterior probabilities can evidence of whether the coefficient is zero or non-zero. Overall, the marginal probability that the coefficient for log(NOX) is non-zero is 0.985, while the posterior probabilities for inclusion of log(HC) and log(SO2) are 0.265 and 0.126 respectively. These are obtained by summing model probabilities over models which include each of the variables, and provide an overall measure for how likely β_j equals zero. Caution should be used in interpreting the marginal posterior probabilities

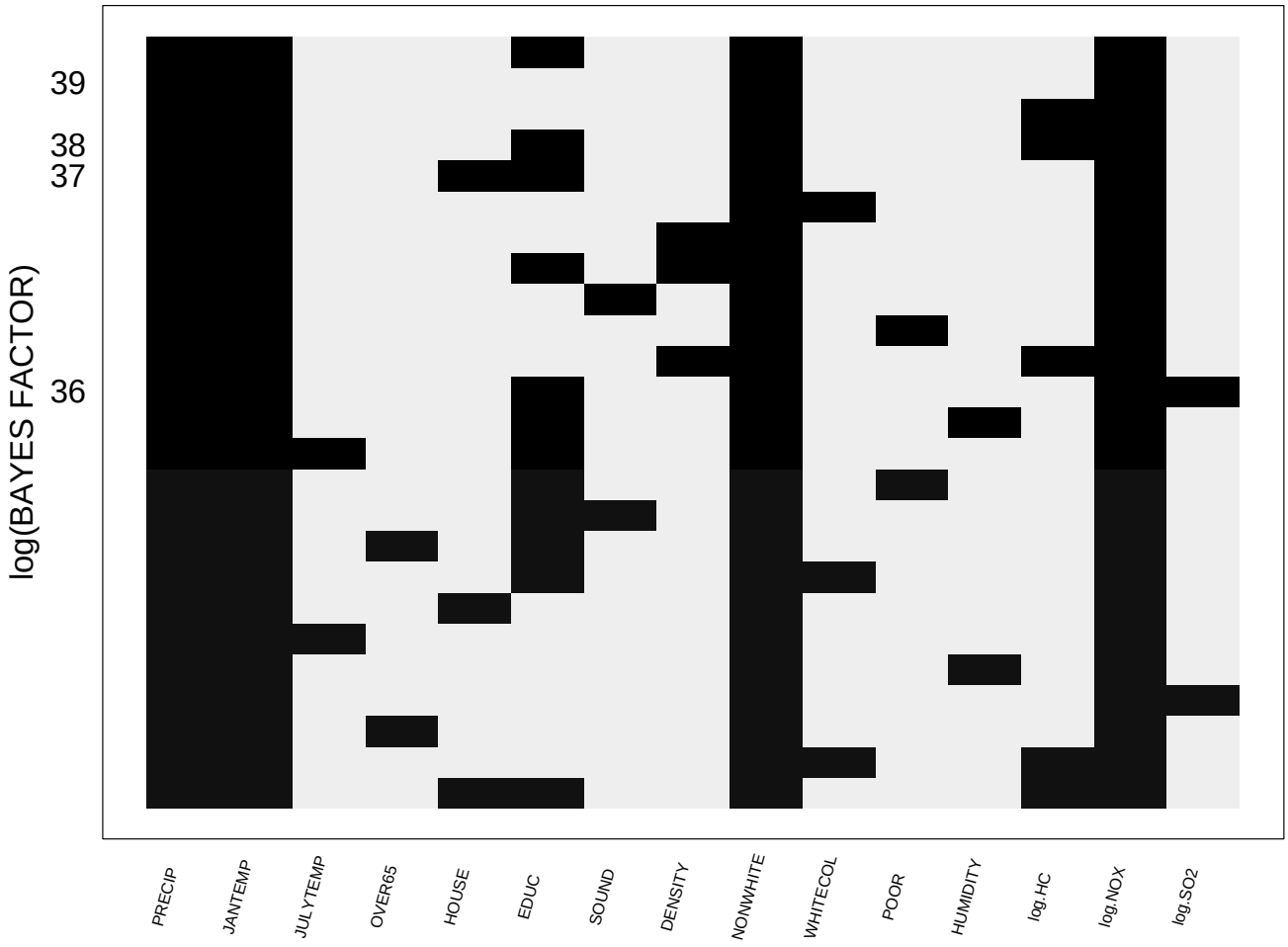

Figure 13.1: Top 25 models based on BIC for the pollution and mortality example.

with many (highly) correlated predictors, as the variables may "split" the posterior mass among themselves and individually receive small (less than 0.5) marginal probabilities of inclusion, while the overall probability that at least one should be included is quite large (the "dilution" effect (Chipman et al., 2001)). In this example, there is overall positive evidence after model averaging for a pollution effect, but the posterior probabilities are spread out over many models.

This analysis only incorporates uncertainty regarding covariates, however, issues of outliers and transformations of variables are also important here. While log transformations had been suggested by other authors, the choice of which transformation to use, outlier detection and accommodation are other aspects of model uncertainty that can by addressed by BMA (Hoeting et al., 1996).

13.7.2 O-ring Failures

On January 27, 1986, engineers that built the space shuttles warned National Aeronautics and Space Administration (NASA) that the *Challenger* should not be launched the next day because of temperatures predicted to be 31^o F and the risk of fuel seal failures at cold temperatures. While the evidence regarding failures seemed inconclusive, the decision was made to proceed with the launch. Statistical models for the number of O-ring failures in each launch as a function of temperature (and pressure) have been explored by Dalal et al. (1989), Lavine (1991), and Draper (1995). The analysis below uses an indicator of at least one O-ring failure in a launch as a response (rather than the number of failures out of the 6 to avoid issues of dependence) with temperature and pressure as potential predictor variables.

Generalized linear models provide a natural extension of linear regression models for binary and other non-normal data. For o-ring failures, it is natural to model the failure indicator Y_i as Bernoulli with probability π_i. As in linear models, we may wish to explore how the mean π_i may depend on other covariates, Xv. If we let the linear combination of the predictors be denoted by $\boldsymbol{\eta}$, ($\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$) then in the normal linear regression, a natural starting point is to assume that the mean $\boldsymbol{\mu}$ is identical to $\boldsymbol{\eta}$, the linear predictor. For binary regression, the mean is a probability and so any posterior distributions on π_i should be constrained to the interval (0,1). While such constraints are

difficult to impose using an "identity link" where the mean π_i is equal to η_i, other functions are more natural for "linking" the mean of binary data to the linear predictor and enforcing the constraints on the probabilities. A widely used choice, the logit link corresponds to modeling the log odds of seal failure as a linear function of the predictors ($\text{logit}(\pi_i) = \log(\pi_i/(1-\pi_i)) = \eta_i$). The probit link function is based on the inverse cdf of the normal distribution, $\Phi^{-1}(\pi_i) = \eta_i$, and is very similar to the logit model except out in the extreme tails (i.e. at low temperatures). The complementary log-log link, $\log(-\log(1-\pi_i))$, is another choice. Figure 13.2 shows the fitted probabilities as a function of temperature for the three link functions. For extrapolation, in the case of prediction at 31^o F, the models may lead to different predictions. Rather than using a single link function, in BMA, the predictions are weighted by how much weight (a posteriori) is given to each link function and incorporates uncertainty regarding the structural form of the model.

Draper (1995) models the number of o-ring failures and considers structural (model) uncertainty in predicting failure at 31^o F due to choice of link function and predictors (temperature, pressure and quadratic terms in temperature). For illustration, we consider the choice of link function (logit, probit, and complementary log-log) and within each link function, the inclusion/exclusion of temperature and pressure (linear terms only) leading to 12 possible models (3 link x 2^2 covariates). The models under the three link functions but with the intercept only, are all reparameterizations of the same model (the null model), so that in reality there are only 10 unique models under consideration. Assignment of uniform model probabilities in the latter case leads to a prior probability that temperature should be included of 0.6 and less weight on the null model (1/10); assigning uniform prior probabilities in the model space with 12 models (realizing that three are equivalent), leads to equal odds a priori that temperature (or pressure) should be included and probability of 1/4 on the null model. While prior odds, other than 1, for temperature (or pressure) inclusion could be based on subjective knowledge, even odds will be used for the rest of the example as a "default" or "reference" analysis.

Using BIC (13.11), the model probabilities can be calculated from deviances in the output of any GLM software. Using prior probabilities $p(\mathcal{M}_k) = 1/12$, the posterior probabilities for each of the 12 models based on choice

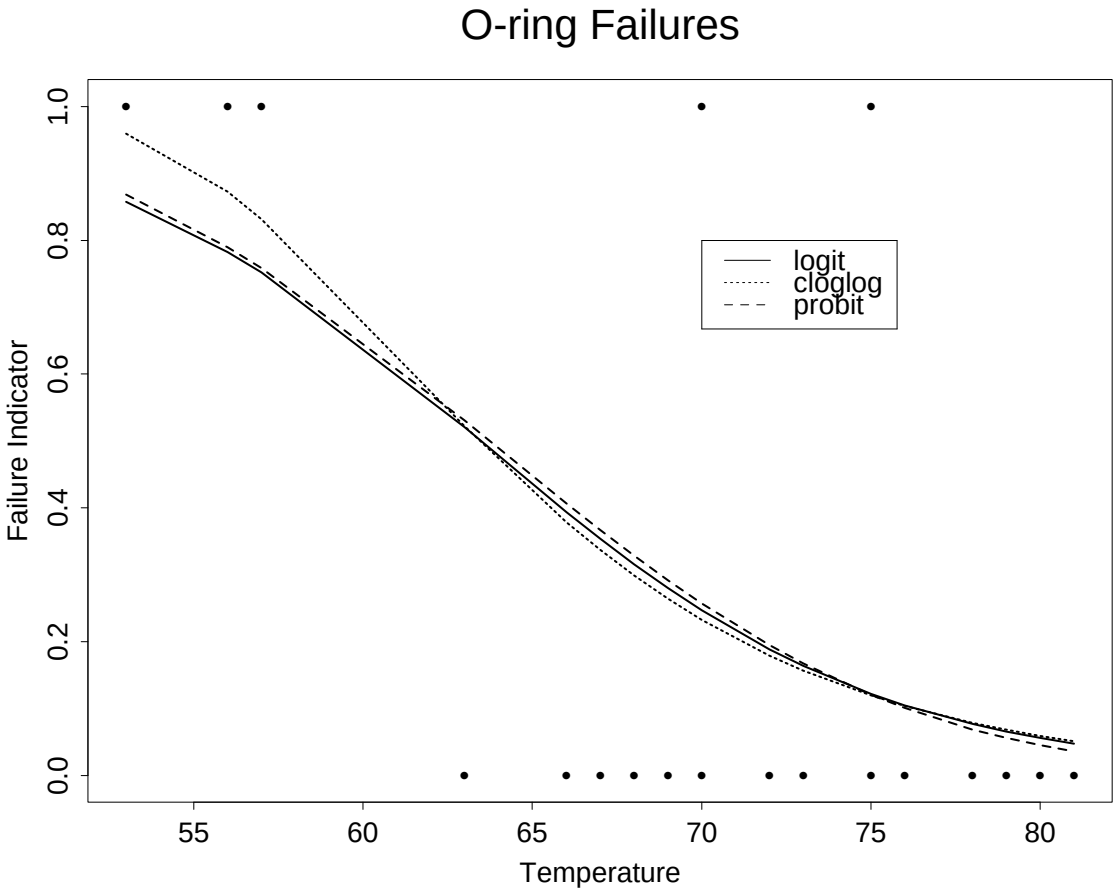

Figure 13.2: Fitted probabilities of O-ring failure using the logit, probit, and complementary log-log link functions.

of link function and covariates are given in Table 13.1. While the prior

	logit	probit	cloglog
Intercept only	0.0167	0.0167	0.0167
Temperature	0.1861	0.1803	0.2753
Pressure	0.0083	0.0082	0.0084
Temperature and Pressure	0.0835	0.0853	0.1143

Table 13.1: Posterior model probabilities based on the BIC approximation for choice of link function and inclusion of temperature and pressure.

probability that temperature is related to failure is 1/2, the posterior probability that temperature is related to failure averaging over link functions is 0.9249, (obtained by adding probabilities of all models that include temperature: 0.1861+0.1803+0.2753+0.0853+0.0835+0.1143) ; similarly the posterior probability that pressure is related to failure is $0.3081 = 0.0083+0.0082+0.0084+0.0835+0.0853+0.1143$.

There is a slightly higher probability in favor of the complementary log-log link, but otherwise the data are not strongly conclusive about link choice, which is not totally surprising given the small differences in probabilities over the range of observed data. For the problem of predicting failure at low temperatures (an extrapolation), the choice of link function can be influential, and in this case model averaging reflects the uncertainty regarding model form. While all models with temperature predict that the probability of failure at 31^o F is above 0.99, the posterior mean with model averaging for the failure probability is 0.92, which incorporates uncertainty about whether temperature has an effect and choice of link function.

Using the normal approximation to the posterior distribution for the coefficients under each model, the posterior distribution and 95% posterior probability intervals for the failure probability under BMA can be obtained by Monte Carlo samples, by first drawing a model with replacement according to the posterior model probabilities, and then drawing the linear predictor $\boldsymbol{\eta}$ based on the asymptotic normal approximation to the posterior distribution for $\boldsymbol{\eta}$ given that model. Applying the inverse link to the linear predictor under the sampled model, leads to one draw from the (approximate) posterior distribution of the probability of failure at 31^o F. This is repeated to provide a Monte Carlo sample from the (approximate) posterior distribution (bottom histogram in Figure 13.3). Using the mixture model under model averaging

and drawing Monte Carlo samples, the 95% posterior probability interval for the failure probability is 0.2568 to 1, which is similar to results in Lavine (1991) and Draper (1995), but slightly wider (the other analyses were based on modeling the number of o-ring failures while here, the indicator of at least one failure was the response). The uncertainty about whether temperature and/or pressure should be included leads to bimodality in the posterior distribution of the failure probability under BMA (bottom plot in Figure 13.3) and a long left tail.

13.8 SUMMARY

In problems where there is a plethora of models and no scientific rational that requires that a single model be used, Bayesian model averaging addresses many of the problems associated with model selection. Model uncertainty is almost always an issue in statistical analysis of data. Bayesian model averaging provides a coherent approach for incorporating uncertainty due to predictor selection, transformations, outliers, model form, and much more. The examples illustrate how BMA can be carried out without using any specialized software using "objective" prior distributions based on BIC. While posterior model probabilities can often be approximated using BIC, there is an extensive literature on other approaches for implementing BMA in a wide variety of applications, using both subjective and objective prior distributions. The articles by Berger and Pericchi (2001), Chipman et al. (2001), Hoeting et al. (1999) and Clyde (1999) provide a description of current practice and open issues in model selection and model averaging, as well as additional history and references for the interested reader.

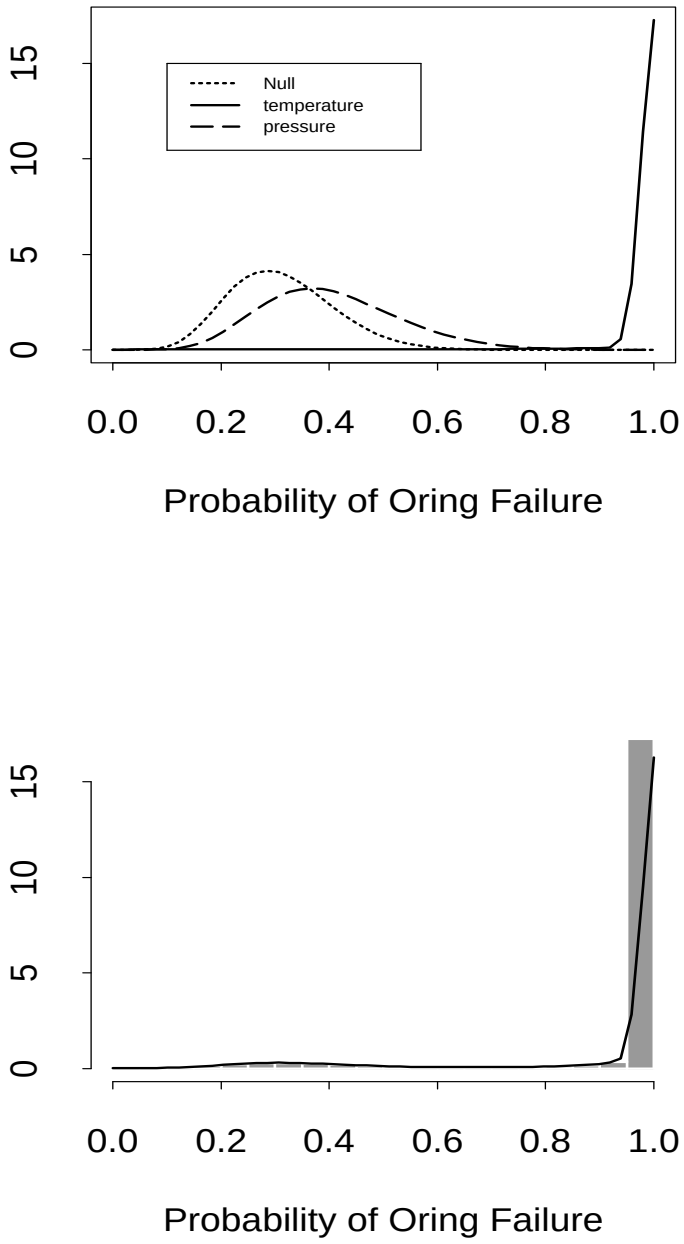

Figure 13.3: Posterior distributions for probability of o-ring failure at 31^o F and pressure = 200. The top plot shows the posterior distributions of the probability of oring failure using the complementary log-log link functions for the null model (intercept only), and models with temperature alone and pressure alone. The smooth curve in the bottom plot shows the posterior distribution of the probability of oring failure under model averaging. The histogram is based on the Monte Carlo sample which was used to construct the density.

EXERCISES

13.1 Using the conjugate prior framework, show that conditional distribution for γ_j given the other γ_k, $k \neq j$ and $\mathbf{Y}$ is Bernoulli. Describe how to implement a Gibbs sampler to explore the posterior distribution of models.

13.2 Consider the linear regression model, with Zellner's g-prior and the non-informative prior on σ^2, $p(\sigma^2) = 1/\sigma^2$.

(a) Find the predictive distribution of a single future observation Y_o at $X = X_o$ under model averaging in the linear regression model.

(b) What is the mean of the predictive distribution?

(c) What is the variance of Y_o in the predictive distribution under model averaging?

(d) Rather than using Monte Carlo sampling, describe how one could construct a 95% prediction interval using the cumulative distribution functions under model averaging.

13.3 Let $\mathcal{M}$ denote the full model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ and consider comparing $\mathcal{M}$ to the null model where $\mathcal{M}^* : \boldsymbol{\beta} = 0$.

(a) Using a noninformative prior distribution for σ^2, $p(\sigma^2) = 1/\sigma^2$, and Zellner's g-prior for $\boldsymbol{\beta}$, $\boldsymbol{\beta} \sim N(0, \sigma^2 g(\mathbf{X}'\mathbf{X})^{-1})$, find the marginal likelihood for $\mathbf{Y}$ under model $\mathcal{M}$.

(b) Find the marginal likelihood for $\mathbf{Y}$ under the model $\mathcal{M}^* : \boldsymbol{\beta} = 0$ (use the same non-informative prior for σ^2).

(c) Using equal prior probabilities on $\mathcal{M}$ and $\mathcal{M}*$, show that the posterior odds of $\mathcal{M}^*$ to $\mathcal{M}$ goes to a non-zero constant $(1+g)^{(k-n)/2}$ as the classical F statistic for testing $\mathcal{M}$ goes to infinity. (Even though one becomes certain that $\mathcal{M}^*$ is wrong, the posterior odds or Bayes Factor does not go to zero, which have led many to reject g-priors for model selection; this criticism applies equally for model averaging - see Berger and Pericchi (2001) for alternatives.)

(d) Using the marginal likelihood under $\mathcal{M}$, find the maximum likelihood estimate of g, $\hat{g}$. Using the "Empirical Bayes" g-prior ($\hat{g}$ substituted for g in the g-prior), what happens to the posterior odds as F goes to infinity?

(e) For model $\mathcal{M}$, consider a Cauchy prior for $\boldsymbol{\beta}|\sigma^2$ (Zellner and Siow, 1980) of the form

$$p(\boldsymbol{\beta}|\sigma^2) = \Gamma\left(\frac{q+1}{2}\right)\frac{|\mathbf{X}'\mathbf{X}|^{1/2}}{(g\sigma^2)^{q/2}}\left(1+\frac{\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}}{g\sigma^2}\right)^{-(q+1)/2}$$

rather than the conjugate normal. Show that the Cauchy distribution can be written as a scale mixture of normal distributions, where $\boldsymbol{\beta}|\sigma^2, \lambda$ is $N(0, \frac{g\sigma^2}{\lambda}(\mathbf{X}'\mathbf{X})^{-1})$ and λ has a gamma$(1/2, 2)$ distribution (parameterized so that the mean is 1).

(f) Investigate the limiting behavior of the posterior odds under the Cauchy prior distribution when F goes to infinity. *Using the scale mixture of normals representation, the posterior odds can be obtained using one-dimensional integration with respect to* λ.

13.4 Consider an improper prior distribution for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ given σ^2: $p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma})$ proportional to $|g\mathbf{X}'_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}}|^{1/2}$.

(a) What happens to the prior distribution of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ if g is multiplied by an arbitrary constant?

(b) Find the posterior distribution for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$. What happens to the posterior distribution for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ if g is multiplied by an arbitrary constant?

(c) What consequence does the use of improper priors have for model selection or model averaging?

(d) Find the posterior distribution of $\boldsymbol{\gamma}$. What happens to the posterior distribution of $\boldsymbol{\gamma}$ when g is multiplied by an arbitrary constant?

(e) If the prior distribution is $p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}) \propto c$, where c does not depend on the data, show that the posterior model probabilities are not invariant to scale changes in $\mathbf{X}$, i.e. if a

variable $\mathbf{X}_j$ is replaced by $a\mathbf{X}_j$ for some non-zero scalar constant a, then the posterior model probabilities depend on a, even though the models using X_j and aX_j are equivalent.

Bibliography

Akaike, H. (1973). "Information theory and an extension of the maximum likelihood principle." In *Second International Symposium on Information Theory*, eds. B. Petrox and F. Caski, 267.

Berger, J. O. and Pericchi, L. R. (2001). "Objective Bayeisan Methods for Model Selection:Introduction and Comparison." In *IMS Lecture Notes - Mongraph Series, Volume 38*, 135– 193.

Chipman, H. (1996). "Bayesian Variable Selection With Related Predictors." *The Canadian Journal of Statistics*, 24, 17–36.

Chipman, H., George, E., and McCulloch, R. (2001). "The Practical Implementation of Bayesian Model Selection." In *IMS Lecture Notes - Mongraph Series, Volume 38*, 65–134.

Clyde, M. (1999). "Bayesian Model Averaging and Model Search Strategies (with discussion)." In *Bayesian Statistics 6*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, 157–185. Oxford: Oxford University Press.

— (2001). "Discussion of "The Practical Implementation of Bayesian Model Selection"." In *IMS Lecture Notes - Monograph Series, Volume 38*, 117–124.

Clyde, M. and George, E. (2000). "Flexible Empirical Bayes Estimation for Wavelets." *Journal of the Royal Statistical Society, Series B*, 62, 681–698.

Dalal, S. R., Fowlkes, E. B., and Hoadley, B. (1989). "Risk Analysis of the Space Shuttle: Pre-Challenger Prediction of Failure." *Journal of the American Statistical Association*, 84, 945–957.

Dawid, A. and Lauritzen, S. (2001). "Compatible Prior Distributions." In *Bayesian Methods with Applications to Science, Policy, and Official Statis-*

tics, Selected papers from ISBA 2000: The sixth world meeting of the international society for Bayesian analysis, ed. E. I. George, 109–118. Eurostat.

Draper, D. (1995). "Assessment and Propagation of Model Uncertainty (Disc: P71-97)." *Journal of the Royal Statistical Society, Series B, Methodological*, 57, 45–70.

Fernández, C., Ley, E., and Steel, M. F. (2001). "Benchmark priors for Bayesian model averaging." *Journal of Econometrics*, 381–427.

Foster, D. P. and George, E. I. (1994). "The Risk Inflation Criterion for Multiple Regression." *The Annals of Statistics*, 22, 1947–1975.

Furnival, G. M. and Wilson, Robert W., J. (1974). "Regression By Leaps and Bounds." *Technometrics*, 16, 499–511.

George, E. I. and Foster, D. P. (2000). "Calibration and empirical Bayes variable selection." To appear in *Biometrika*.

George, E. I. and McCulloch, R. E. (1993). "Variable Selection Via Gibbs Sampling." *Journal of the American Statistical Association*, 88, 881–889.

— (1997). "Approaches for Bayesian Variable Selection." *Statistica Sinica*, 7, 339–374.

Hodges, J. S. (1987). "Uncertainty, Policy Analysis and Statistics (C/R: P276-291)." *Statistical Science*, 2, 259–275.

Hoeting, H. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). "Bayesian model averaging: A tutorial (with discussion)." *Statistical Science*, 14, 382–417.

Hoeting, J., Raftery, A. E., and Madigan, D. (1996). "A Method for Simultaneous Variable Selection and Outlier Identification in Linear Regression." *Computational Statistics and Data Analysis*, 22, 251–270.

Kass, R. E. and Raftery, A. E. (1995). "Bayes Factors." *Journal of the American Statistical Association*, 90, 773–795.

Lavine, M. (1991). "Problems in Extrapolation Illustrated With Space Shuttle O-ring Data (Com: P921-922)." *Journal of the American Statistical Association*, 86, 919–921.

Lavine, M. and Schervish, M. J. (1999). "Bayes Factors: What They Are and What They Are Not." *The American Statistician*, 53, 119–122.

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). "Bayesian Model Averaging for Linear Regression Models." *Journal of the American Statistical Association*, 92, 179–191.

Ramsey, F. and Schafer, D. (2002). *The Statistical Sleuth: A Course in Methods of Data Analysis*. Duxbury Press.

Schwarz, G. (1978). "Estimating the Dimension of a Model." *The Annals of Statistics*, 6, 461–464.

Smith, M. and Kohn, R. (1996). "Nonparametric Regression Using Bayesian Variable Selection." *Journal of Econometrics*, 75, 317–343.

Weisberg, S. (1985). *Applied Linear Regression (Second Edition)*. Wiley.

Zellner, A. (1986). "On Assessing Prior Distributions and Bayesian Regression Analysis With g-prior Distributions." In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, 233–243. North-Holland/Elsevier (Amsterdam; New York).

Zellner, A. and Siow, A. (1980). "Posterior Odds Ratios for Selected Regression Hypotheses." In *Bayesian Statistics: Proceedings of the First International Meeting held in Valencia (Spain)*, 585–603. University of Valencia (Spain).