# A Bayesian Sampling Approach to Regression Model Checking

Iain Pardoe *
School of Statistics, University of Minnesota, Minneapolis, MN 55455
i.pardoe@stat.umn.edu

May 28, 2000

**Abstract**

A necessary step in any regression analysis is checking the fit of the model to the data. Graphical methods are often employed to allow visualization of features that the data should exhibit if the model was to hold. Judging whether such features are present or absent in any particular diagnostic plot can be problematic. In this article I take a Bayesian approach to aid in this task. The "unusualness" of some data with respect to a model can be assessed using the predictive distribution of the data under the model; an alternative is to use the posterior predictive distribution. Both approaches can be given a sampling interpretation that can then be utilized to enhance regression diagnostic plots such as marginal model plots.

**Key Words:** Model criticism; Predictive distribution; Graphical assessment; Diagnostic plot; Marginal model plot.

*Iain Pardoe is Graduate Student at the School of Statistics, University of Minnesota, Minneapolis, MN 55455 (E-mail: i.pardoe@stat.umn.edu).

# 1 Introduction

Consider a regression problem with $n$ observations of a univariate response $y$ and $p$ predictors, $\boldsymbol{x} = (x_1, \ldots, x_p)^T$. Suppose a model has been derived for the conditional distribution of $y$ given $\boldsymbol{x}$, $\mathrm{F}(y|\boldsymbol{x})$, either from theoretical considerations or data analysis; denote this by $\mathrm{M}(y|\boldsymbol{x}, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of unknown parameters. Assume $\boldsymbol{\theta}$ can be consistently estimated with $\widehat{\boldsymbol{\theta}}$. Before using $\widehat{\mathrm{M}}(y|\boldsymbol{x}) = \mathrm{M}(y|\boldsymbol{x}, \widehat{\boldsymbol{\theta}})$ to address a practical issue, I need to be confident that $\mathrm{M}(y|\boldsymbol{x}, \boldsymbol{\theta})$ provides a *sufficiently accurate* approximation to $\mathrm{F}(y|\boldsymbol{x})$, where the accuracy is gauged relative to the practical issue. In other words, acknowledging the insight of Box (1979) that "all models are wrong, but some are useful", how can I assess if $\mathrm{M}(y|\boldsymbol{x}, \boldsymbol{\theta})$ is useful? I propose a new application of the *Bayesian sampling approach* to that question in this article. I give a rationale for the methodology in Section 2 and describe my graphical application of it in Section 3. I provide details for normal linear models in Section 4 and additive models in Section 5. Section 6 contains a discussion.

# 2 Rationale

## 2.1 Bayesian model checking diagnostics

To introduce ideas and keep notation concise, consider assessing how well a model $\mathrm{M} = \mathrm{M}(y|\boldsymbol{\theta})$ fits some potential data $\boldsymbol{y} = (y_1, \ldots, y_n)^T$, where $\boldsymbol{\theta}$ is assumed to have a prior probability distribution. Box (1980) proposed a Bayesian diagnostic for checking $\mathrm{M}$ based on the following. Conditional on $\mathrm{M}$, the marginal, or predictive, distribution of $\boldsymbol{y}$ can be described by its density

$$f(\boldsymbol{y}|\mathrm{M}) = \int f(\boldsymbol{y}|\boldsymbol{\theta}, \mathrm{M}) f(\boldsymbol{\theta}|\mathrm{M}) \, \mathrm{d}\boldsymbol{\theta} \tag{1}$$

where $f(\boldsymbol{y}|\boldsymbol{\theta}, \mathrm{M})$ is the likelihood for $\boldsymbol{y}$ and $f(\boldsymbol{\theta}|\mathrm{M})$ is the prior density of $\boldsymbol{\theta}$. Once actual data $\boldsymbol{y}_d$ are available, $\mathrm{M}$ can be assessed by referring the value of the predictive density at $\boldsymbol{y}_d$, $f(\boldsymbol{y}_d|\mathrm{M})$, to the density function $f(\boldsymbol{y}|\mathrm{M})$. One way to do this is to calculate

$$\alpha = \Pr(f(\boldsymbol{y}|\mathrm{M}) < f(\boldsymbol{y}_d|\mathrm{M})) \tag{2}$$

where the probability is calculated under $\mathrm{M}$. A "small" value of $\alpha$ indicates that $\boldsymbol{y}_d$ would be unlikely to be generated by $\mathrm{M}$, and thus calls $\mathrm{M}$ into question. More generally, $\mathrm{M}$ can be assessed by referring the value of the predictive density of some relevant checking function, $g_i(\boldsymbol{y})$, at $\boldsymbol{y}_d$ to its predictive density, for a variety of $g_i$. Examples of useful $g_i$ in practice include residuals, order statistics, and moment estimators.

Box's approach can only be used with proper priors, since otherwise (1) does not exist. Rubin (1984) proposed an alternative approach that does not require proper priors, using the posterior predictive density

$$f(\boldsymbol{y}|\boldsymbol{y}_d, \mathrm{M}) = \int f(\boldsymbol{y}|\boldsymbol{\theta}, \mathrm{M}) f(\boldsymbol{\theta}|\boldsymbol{y}_d, \mathrm{M}) \, \mathrm{d}\boldsymbol{\theta}$$

where $f(\boldsymbol{\theta}|\boldsymbol{y}_d, \mathrm{M})$ is the posterior density of $\boldsymbol{\theta}$. Again, diagnostics similar to (2) and checking functions $g_i$ can be constructed. Rubin's approach has been extended by Gelman, Meng, and Stern (1996).

## 2.2 A sampling interpretation

Another way to think about Box's approach is in terms of a sampling simulation. Gelman, Meng, and Stern (1996) provide references to many papers that discuss this interpretation. The idea is to draw a value of $\boldsymbol{\theta}$ from its prior distribution, and then generate a sample of $n$ realizations from the model $\mathrm{M}$ indexed by this $\boldsymbol{\theta}$. Repeat this process a large number $m$ of times and then compare the data $\boldsymbol{y}_d$ to the $m$ realizations from $\mathrm{M}$. Then, intuitively, if $\boldsymbol{y}_d$ "looks like" a typical realization from $\mathrm{M}$, there is no reason to doubt the fit of $\mathrm{M}$. On the other hand, if $\boldsymbol{y}_d$ appears to be very "unusual" with respect to the $m$ realizations from $\mathrm{M}$, then $\mathrm{M}$ is called into question. To do this in practice, methods for comparing $\boldsymbol{y}_d$ to the $m$ realizations from $\mathrm{M}$ and measures of "unusualness" need to be developed. But once done, the methodology can be applied in any situation where samples can be generated from the prior distribution for $\boldsymbol{\theta}$. In particular, the methodology can be applied in situations where quantities such as (2) cannot be derived analytically.

Rubin's approach can also be cast in sampling terms, with each value of $\boldsymbol{\theta}$ drawn from its posterior rather than prior distribution. The crucial difference between the two approaches can therefore be considered in terms of the choice of "sampling distribution" for $\boldsymbol{\theta}$ used to generate realizations from $\mathrm{M}$. This indicates that the two approaches are attempting to address slightly different questions:

- Box considers if the data are consistent with a family of models indexed by parameters whose variability is modeled only by prior beliefs. For example, (2) contrasts information from the *prior* and data, and checks their compatibility.
- Rubin considers if the data are consistent with a particular model *that has been fit to the data* and which is indexed by parameters whose variability depends on both data and prior beliefs. Diagnostics similar to (2) contrast information from the *posterior* and data, and check their compatibility.

Thus, the two approaches differ on how $\boldsymbol{\theta}$ and the data are being viewed. If, when $\mathrm{M}$ holds for the data, the information in the data is to be used to update belief about $\boldsymbol{\theta}$ (ie, to obtain its posterior distribution), Rubin's approach is more appropriate. If not, and belief about $\boldsymbol{\theta}$ is to remain the same regardless of the information in the data, Box's approach ought to be used. Box's approach would therefore be used in a situation where prior knowledge is *well-established*. This does not necessarily correspond to a prior distribution with small variance. Box (1980) described the problem of assessing a normal model for estimating the mean $\mu$ for a single batch of manufactured items, where $\mu$ is assumed to arise from a normally distributed industrial process with mean $\mu_0$ and variance $\sigma_0^2$. Knowledge about the industrial process could be well-established but the prior distribution for $\mu$ might have large variance (ie large batch-to-batch variation).

Nevertheless, as prior information becomes more vague, it seems unlikely that information from the data would *not* be used to update belief about $\boldsymbol{\theta}$ (that is, when $\mathrm{M}$ holds and the data have *relevant* information about $\boldsymbol{\theta}$). With Box's approach, any particular $\boldsymbol{y}_d$ will become less likely to call $\mathrm{M}$ into question, as prior information, and hence the predictive distribution of $\boldsymbol{y}$, becomes more vague. As mentioned earlier, in the limit with an improper prior, Box's approach cannot even be used.

Some aspects of the model being checked may depend only on $\boldsymbol{\theta}$ itself, rather than on a sample of $n$ realizations from the model $\mathrm{M}(\boldsymbol{\theta})$. For example, $\boldsymbol{\theta}$ might represent predicted values for $\boldsymbol{y}$ under $\mathrm{M}$. If *model-free* predicted values were available, these could be compared directly with the $\boldsymbol{\theta}$ samples to assess the fit of the model.

# 3 A graphical application

## 3.1 Marginal model plots

I now return to the regression setting of Section 1 to apply the ideas in Section 2 to a particular graphical method for comparing $\mathrm{F}(y|\boldsymbol{x})$ to $\widehat{\mathrm{M}}(y|\boldsymbol{x})$. Following on from Cook and Weisberg (1997), $\mathrm{F}(y|\boldsymbol{x}) = \mathrm{M}(y|\boldsymbol{x})$ for all values of $\boldsymbol{x}$ in its sample space if and only if $\mathrm{F}(y|h) = \mathrm{M}(y|h)$ for all functions $h = h(\boldsymbol{x})$. So, a comparison between $\mathrm{F}(y|\boldsymbol{x})$ and $\widehat{\mathrm{M}}(y|\boldsymbol{x})$ can be made by comparing characteristics of $\mathrm{F}(y|h)$ and $\widehat{\mathrm{M}}(y|h)$ for various $h$. Particular characteristics that can be useful to compare include mean and variance functions.

To compare mean functions for example, plot $y$ versus $h$ for a particular $h$. Add a non-parametric mean estimate, say a cubic smoothing spline with fixed smoothing parameter, to the plot; denote this by $\widehat{\mathrm{E}}_{\mathrm{F}}(y|h)$, where $\mathrm{E}_{\mathrm{F}}$ denotes expectation under $\mathrm{F}$. The corresponding mean estimate under $\widehat{\mathrm{M}}$ is $\widehat{\mathrm{E}}_{\widehat{\mathrm{M}}}(y|h)$, where $\mathrm{E}_{\widehat{\mathrm{M}}}$ denotes expectation under $\widehat{\mathrm{M}}$. Since $\mathrm{E}_{\widehat{\mathrm{M}}}(y|h) = \mathrm{E}[\mathrm{E}_{\widehat{\mathrm{M}}}(y|\boldsymbol{x})|h]$, $\widehat{\mathrm{E}}_{\widehat{\mathrm{M}}}(y|h)$ can be obtained from a non-parametric mean estimate for the regression of the fitted values under $\widehat{\mathrm{M}}$, $\mathrm{E}_{\widehat{\mathrm{M}}}(y|\boldsymbol{x})$, on $h$. Add $\widehat{\mathrm{E}}_{\widehat{\mathrm{M}}}(y|h)$ to the plot with $\widehat{\mathrm{E}}_{\mathrm{F}}(y|h)$ to obtain a *marginal model plot* (MMP) for the mean in the (marginal) direction $h$. Using the same method and smoothing parameter for the mean estimates under $\widehat{\mathrm{M}}$ and $\mathrm{F}$ allows point-wise comparison of the two estimates, since any estimation bias should cancel (Bowman and Young 1996).

Ideas for selecting useful functions $h$ to consider in practice are given in Cook and Weisberg (1997), and include fitted values, individual predictors, and linear combinations of the predictors. Further discussion of this issue is prominent in the *dimension-reduction* literature, and work is in progress to develop complementary techniques in this context. Some examples include *principal Hessian directions*, due to Li (1992), and *sliced average variance estimation*, due to Cook and Weisberg (1991), which can often find functions $h$ where lack of fit is most likely to be observed. If $\mathrm{M}$ is an accurate approximation to $\mathrm{F}$, then for any quantity $h$ the marginal mean estimates should agree, $\widehat{\mathrm{E}}_{\mathrm{F}}(y|h) \approx \widehat{\mathrm{E}}_{\widehat{\mathrm{M}}}(y|h)$. Any indication that the estimated marginal means do not agree for one particular $h$ calls $\mathrm{M}$ into question; if they agree for a variety of plots, there is support for $\mathrm{M}$.

### 3.2 Bayes marginal model plots

A problem that arises with using MMP's in practice is deciding, relative to the variation in the data, when the estimated marginal means agree and when they do not agree. How large do discrepancies between the estimated marginal means have to be to call $\mathrm{M}$ into question? Porzio and Weisberg (1999) provide frequentist methodology to address this issue: point-wise reference bands to aid visualization and statistics to calibrate discrepancies. An alternative approach is to apply the ideas discussed in Section 2.

Even if $\mathrm{M}(y|\boldsymbol{x},\boldsymbol{\theta}) = \mathrm{F}(y|\boldsymbol{x})$, the estimated marginal means in a MMP would not match exactly. So, a technique is needed to visualize the variability in $\mathrm{M}$ to assess whether it would be reasonable for the data to be generated by such an $\mathrm{M}$. The sampling interpretation for the Bayesian model checking diagnostics of Box and Rubin provides such a technique: for any particular MMP, just calculate mean estimates for fitted values corresponding to individual samples from either the prior distribution (under Box's approach) or posterior distribution (under Rubin's approach) of $\boldsymbol{\theta}$. Then, instead of adding the mean estimate under $\widehat{\mathrm{M}}$ to the plot of the mean estimate under F, add a mean estimate for each sample from $\widehat{\mathrm{M}}$, and obtain what I call a *Bayes marginal model plot* (BMMP) for the mean. These plots were called *Gibbs* marginal model plots in Cook and Pardoe (2000). The $\boldsymbol{\theta}$ samples are being used directly here, as suggested in the final paragraph of Section 2.2.

If enough samples are taken, say $m = 100$, the Bayes mean estimates will form a mean estimate *band* under $\widehat{\mathrm{M}}$. The plot then provides a visual way of determining whether there is any evidence to contradict the possibility that $\mathrm{F}(y|h) = \mathrm{M}(y|h)$. If, for a particular $h$, the mean estimate under F lies substantially outside the mean estimate band under $\widehat{\mathrm{M}}$, then $\mathrm{M}$ is called into question. If, no matter what the function $h$ is, the mean estimate under F lies broadly inside the mean estimate band under $\widehat{\mathrm{M}}$, then perhaps $\mathrm{M}$ provides an accurate description of the conditional distribution of $y|\boldsymbol{x}$.

## 4 Normal Linear Models

The normal linear regression model can be written

$$y_i|\boldsymbol{x}_i = \mathrm{E}(y|\boldsymbol{x}_i) + e_i/\sqrt{w_i}, \quad i = 1, \ldots, n$$

where $\mathrm{E}(y|\boldsymbol{x}_i) = \boldsymbol{\beta}^T\boldsymbol{x}_i$, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters, $\boldsymbol{x}_i$ is the $p \times 1$ vector of predictor values for the $i$-th observation, the errors $e_i$ are normally distributed with mean 0 and variance $\sigma^2$, and the weights $w_i > 0$ are known, positive numbers. An intercept term can be included within this framework by setting one of the predictors equal to a constant. Defining $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$, $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \sigma^2)^T$, and $\boldsymbol{W} = \mathrm{diag}(w_i)$, then (suppressing the notation for conditioning on $\mathrm{M}$ for clarity)

$$\boldsymbol{y}|(\boldsymbol{X},\boldsymbol{\theta}) \sim \mathrm{N}\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{W}^{-1}\right)$$

The usual non-informative prior for the normal linear regression model is $f(\boldsymbol{\theta}) \propto \sigma^{-2}$. Consider constructing BMMP's in this situation. Since the prior is improper, only Rubin's approach is appropriate. Sampling from the posterior is straightforward since $f(\boldsymbol{\beta},\sigma^2|(\boldsymbol{X},\boldsymbol{y}_d)) = f(\boldsymbol{\beta}|(\boldsymbol{X},\boldsymbol{y}_d,\sigma^2)\, f(\sigma^2|(\boldsymbol{X},\boldsymbol{y}_d))$. In particular, draw a value of $\sigma^2$ from

$$\sigma^2|(\boldsymbol{X},\boldsymbol{y}_d) \sim \mathrm{RSS}\chi^{-2}_{n-p}$$

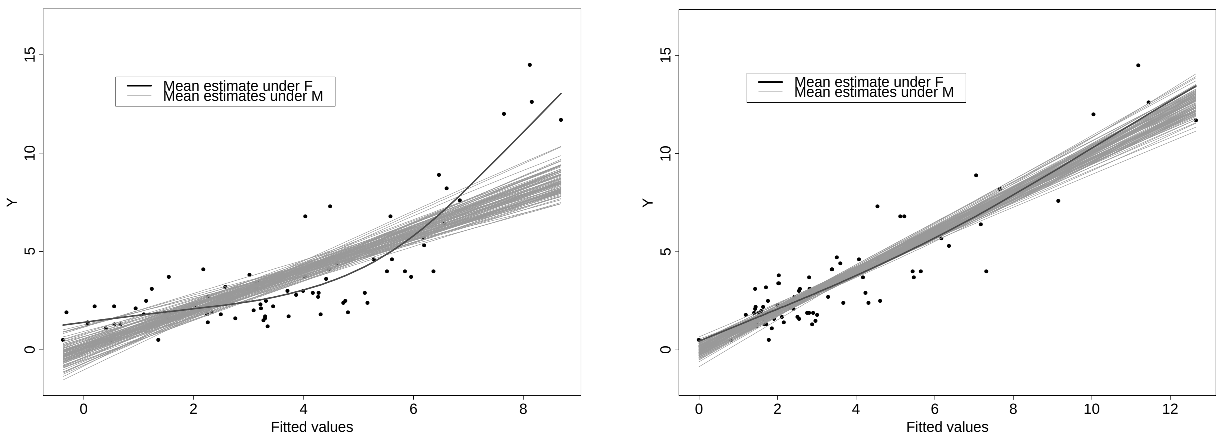
where $\mathrm{RSS}\chi^{-2}_{n-p}$ is the usual weighted residual sum of squares divided by a $\chi^2$ random variable with $n-p$ degrees of freedom. Then, holding $\sigma^2$ fixed, draw a value of $\boldsymbol{\beta}$ from

$$\boldsymbol{\beta}|(\boldsymbol{X},\boldsymbol{y}_d,\sigma^2) \sim \mathrm{N}\left(\widehat{\boldsymbol{\beta}}, \sigma^2(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\right)$$

where $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{y}$ is the usual weighted least squares estimate for $\boldsymbol{\beta}$. The fitted values corresponding to the posterior samples $\boldsymbol{\beta}_t$ are $\boldsymbol{X}\boldsymbol{\beta}_t$, $t = 1, \ldots, m$.

This procedure is straightforward to program. I developed S-PLUS functions to generate the appropriate posterior samples from a linear model object, and to construct corresponding BMMP's for the mean in any user-specified direction $h$. I use the `smooth.spline` function (Hastie and Tibshirani 1990) to obtain the non-parametric mean estimates with user-specified smoothing parameter, $\gamma$. Recall that the smoothing parameters for all smooths in a particular BMMP need to be equal to allow their point-wise comparison. Therefore it is desirable to select $\gamma$ so that the smooths are flexible enough to capture the systematic trends in all the corresponding scatterplots, while not over-fitting too much in any one scatterplot. This is clearly impractical, so a prudent compromise is to graphically select $\gamma$ to capture the systematic trends only in the scatterplots for the data and for $\boldsymbol{X}\widehat{\boldsymbol{\beta}}$. An alternative is to select $\gamma$ automatically, for example by using the most flexible of the smoothing parameters chosen by cross-validation for a subset of the corresponding scatterplots. Given the sometimes erratic behavior of automatic smoothing parameter selection methods, I prefer to use the graphical approach using the plots of the data and $\boldsymbol{X}\widehat{\boldsymbol{\beta}}$ in practice.

Consider an example on the effects of three variables, $AN = \log_{10}(\text{air to naphthalene ratio})$, $Ctime = \log_{10}(\text{contact time})$, and $Btemp = 0.01(\text{bed temperature} - 330)$, on $Y$ = percentage mole conversion of naphthalene to naphthoquinone. The data arose from a chemical experiment in the 1950's to develop a catalyst in the vapor phase oxidation of naphthalene. The variable transformations used result in approximate trivariate normality and appear to stabilize the yield surface. Box and Draper (1987) based an analysis of the data on a full second-order response surface model. BMMP's for the mean in the direction of the fitted values for both a first-order linear model and the full second-order linear model are shown in Figure 1. I selected $\gamma$ so that the smoothing splines had four *effective degrees of freedom* (Hastie and Tibshirani 1990) in both plots.




First-order linear model

Full second-order linear model

Figure 1: BMMP's for the mean in the direction of the fitted values for the naphthalene data.

The BMMP for the first-order model shows the mean estimate under F lying mostly outside the mean estimate band under $\widehat{\mathrm{M}}$. This provides clear evidence to call this model into question. The BMMP for the second-order model shows the mean estimate under F lying inside the mean estimate band under $\widehat{\mathrm{M}}$. There is little evidence in *this* plot to call this model into question. In fact, BMMP's for the second-order model in a variety of directions $h$ all appear to have this characteristic. So, using this graphical diagnostic technique, there appears to be no compelling evidence to question the second-order model. I will return to this example after discussing a method for constructing BMMP's for additive models.

# 5 Additive Models

The additive regression model can be written

$$y_i|\boldsymbol{x}_i = \alpha + \sum_{j=1}^{p} f_j(x_{ij}) + e_i, \quad i = 1, \ldots, n$$

where $f_j$ is a "smooth" function for the $j$-th predictor, $x_{ij}$ is the $i$-th observation of the $j$-th predictor, and the errors $e_i$ are normally distributed with mean 0 and variance $\sigma^2$. Defining $\boldsymbol{\theta} = (\alpha, f_1, \ldots, f_p, \sigma^2)^T$, then

$$\boldsymbol{y}|(\boldsymbol{X}, \boldsymbol{\theta}) = \alpha \boldsymbol{J}_n + \sum_{j=1}^{p} \boldsymbol{f}_j + \boldsymbol{e} \tag{3}$$

where $\boldsymbol{J}_n$ is an n $\times$ 1 vector of ones, $\boldsymbol{f}_j = (f_j(x_{1j}), \ldots, f_j(x_{nj}))^T$, and $\boldsymbol{e}$ is an $n \times 1$ vector of errors.

Hastie and Tibshirani (1990) describe a Bayesian characterization of (3) based on partially improper normal process priors for each $\boldsymbol{f}_j$

$$\boldsymbol{f}_j \sim \mathrm{N}(\mathbf{0}_n, \tau_j^2 \boldsymbol{K}_j^-)$$

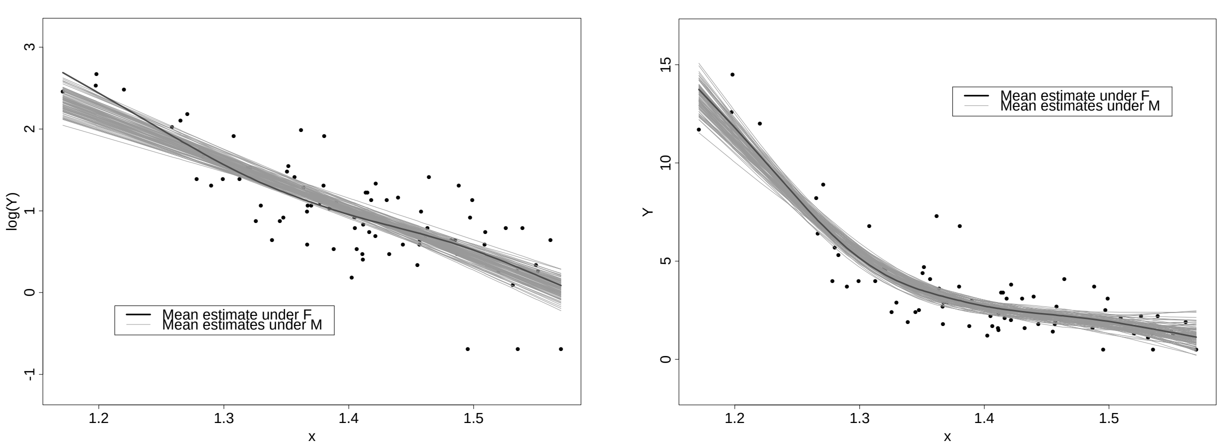
where $\mathbf{0}_n$ is an n $\times$ 1 vector of zeros, and $\boldsymbol{K}_j^-$ is a generalized inverse of a matrix $\boldsymbol{K}_j$ which is related to the construction of the estimate of $\boldsymbol{f}_j$. For example, when $\boldsymbol{f}_j$ is estimated using a symmetric smoother matrix $\boldsymbol{S}_j$ with smoothing parameter $\lambda_j$, then $\tau_j^2 = \sigma^2/\lambda_j$, and $\boldsymbol{K}_j^- = \lambda_j(\boldsymbol{S}_j^- - \boldsymbol{I}_n)^-$, where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix.

Consider constructing BMMP's in this situation. Rubin's approach is appropriate here if most of the information about $\boldsymbol{\theta}$ is going to come from the data rather than the prior. Also, the prior for $\boldsymbol{\theta}$ can be improper if any of the $\boldsymbol{f}_j$ correspond to fixed linear effects, thus necessitating Rubin's approach. Hastie and Tibshirani (2000) derive a way to sample from the posterior of $\boldsymbol{\theta}$ using a stochastic generalization of the backfitting algorithm based on Gibbs sampling. In particular, $\boldsymbol{f}_j$ has posterior

$$\boldsymbol{f}_j | (\boldsymbol{X}, \boldsymbol{y}_d) \sim \mathrm{N}(\boldsymbol{S}_j \boldsymbol{y}_d, \sigma^2 \boldsymbol{S}_j)$$

Then, with $\sigma^2$ and each $\tau_j^2$ held fixed, posterior samples are generated by adding noise $\sigma \boldsymbol{S}_j^{1/2} \boldsymbol{z}$, where $\boldsymbol{z}$ is an $n \times 1$ vector of standard normal random variables, to the partial residual smooths in the backfitting algorithm. If $\sigma^2$ and each $\tau_j^2$ are not held fixed, conjugate inverse gamma priors lead to inverse gamma conditional sampling steps for $\sigma^2$ and each $\tau_j^2$ in the algorithm. The fitted values corresponding to the posterior samples $(\alpha, \boldsymbol{f}_1, \ldots, \boldsymbol{f}_p)_t$ are $\alpha_t \boldsymbol{J}_n + \boldsymbol{f}_{1t} + \cdots + \boldsymbol{f}_{pt}$, $t = 1, \ldots, m$. Hastie and Tibshirani developed S-PLUS functions to carry out their sampling procedure. The code discussed in Section 4 can then be used to construct corresponding BMMP's for the mean in any user specified direction $h$.

Consider the naphthalene data again. The full second-order linear model appears to provide a good fit to the data, but other regression diagnostics suggest some deficiencies. In particular, although all the coefficient estimates are highly significant, three observations with large Cook's distances have a relatively strong influence on some of these estimates. Also, there are hints of dependence in some residual plots. Cook (1998) suggests that the dependence of $Y$ on the predictors is through the single linear combination $x = 0.397AN + 0.445Ctime + 0.802Btemp$. Applying Box-Cox response transformation methodology, a linear model with transformed response $\log(Y)$ and single predictor $x$ may provide an improvement on the second-order model. Alternatively, an additive model with response $Y$ and predictor $x$ may prove more accurate. BMMP's for the mean in the direction of $x$ for both the linear and additive models are shown in Figure 2. I selected $\gamma$ so that the smoothing splines had four effective degrees of freedom in both plots. The posterior sampling for the additive model was carried out with $\sigma^2$ fixed at the unbiased estimate from the additive model fit and the effective degrees of freedom for the smooth for $\boldsymbol{f}$ fixed at four.




Linear model with transformed response

Additive model

Figure 2: BMMP's for the mean in the direction of $x$ for the naphthalene data.

The BMMP for the linear model with transformed response and predictor $x$ has a similar flavor to the BMMP for the second-order model in Figure 1, although the mean estimate under F lies a little closer to the edges of the mean estimate band under $\widehat{\mathrm{M}}$. This feature was similar for other directions $h$ also. This indicates that the linear model with transformed response provides a fit that appears to be almost as good as the fit provided by the second-order model, and, since it is a much simpler model, it might be preferred on parsimonious grounds. As with the second-order model however, further regression diagnostics indicate that this model too has hints of dependency in some residual plots. The BMMP for the additive model shows the mean estimate under F lying well inside the mean estimate band under $\widehat{\mathrm{M}}$. There is little evidence in *this* plot, or indeed in other BMMP's for the mean in alternative directions $h$, to call this model into question. Further regression diagnostics indicate that this model provides a good fit to the data with no clear deficiencies.

# 6 Discussion

BMMP's for the mean offer a quick and easy way to check models graphically. The sampling needs to be done only once for each model and cycling through BMMP's in a variety of directions $h$ provides guidance on the fit of the model. The methodology can be extended to variance function estimates to provide further ways for checking models. MMP's can easily be constructed for other types of regression model such as generalized linear models; using techniques such as *Markov chain Monte Carlo* to sample from the posteriors of the model parameters allows BMMP's to also be constructed for such models.

The examples I considered for normal linear models and additive models adopted Rubin's approach using posterior sampling. An example where Box's approach using prior sampling might be more appropriate is in a situation where follow-up data were being analyzed for a model that has been fit to previous data.

There are other plots used in the area of regression diagnostics that can be difficult to assess relative to the variation in the data. Examples include: residual plots; CERES plots, which are a generalization of partial residual plots and were introduced by Cook (1993); net-effect plots, which aid in assessing the contribution of a selected predictor to a regression and were introduced by Cook (1995). The ideas discussed above would appear to have a rôle to play in the analysis of such plots. Work is in progress on these issues, as well as on developing supplementary Bayesian inference methodology.

Further work is also in progress on making the graphical technique described in this article more precise. In particular, deciding whether the mean estimate under F lies substantially outside or broadly inside the mean estimate band under $\widehat{\mathrm{M}}$ can be a highly subjective process. If it is decided that the F estimate is outside the $\widehat{\mathrm{M}}$ estimate band, the nature of the discrepancy may also provide some guidance on how a better model might be developed. For example, a BMMP may exhibit clear discrepancy as a result of just a few observations in the data—these observations may need to be treated differently to the remainder of the data in drawing overall conclusions.

## Acknowledgments

This research was partially supported by National Science Foundation grant DMS-9703777. The author thanks Trevor Hastie and Rob Tibshirani for help in implementing their S-PLUS code for *Bayesian backfitting*, and the editor, an associate editor, and an anonymous reviewer for helpful comments on an earlier version of this article. S-PLUS code to implement the methods described in this article is available from the author.

# References

Bowman, A. W. and S. Young (1996). Graphical comparison of nonparametric curves. *Applied Statistics 45*, 83–98.

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In R. Launer and G. Wilkinson (Eds.), *Robustness in Statistics*. New York: Academic Press.

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness (with discussion). *Journal of the Royal Statistical Society, Series A (General) 143*, 383–430.

Box, G. E. P. and N. Draper (1987). *Empirical Model-Building and Response Surfaces*. New York: Wiley.

Cook, R. D. (1993). Exploring partial residual plots. *Technometrics 35*, 351–362.

Cook, R. D. (1995). Graphics for studying net effects of regression predictors. *Statistica Sinica 5*, 689–708.

Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions through Graphics*. New York: Wiley.

Cook, R. D. and I. Pardoe (2000). Comment on “Bayesian backfitting” by T. J. Hastie and R. J. Tibshirani. *Statistical Science 15*, 213–216.

Cook, R. D. and S. Weisberg (1991). Comment on “Sliced inverse regression for dimension reduction” by K.-C. Li. *Journal of the American Statistical Association 86*, 316–342.

Cook, R. D. and S. Weisberg (1997). Graphics for assessing the adequacy of regression models. *Journal of the American Statistical Association 92*, 490–499.

Gelman, A., X.-L. Meng, and H. Stern (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica 6*, 733–807.

Hastie, T. J. and R. J. Tibshirani (1990). *Generalized Additive Models*. Boca Raton, FL: Chapman & Hall/CRC.

Hastie, T. J. and R. J. Tibshirani (2000). Bayesian backfitting (with discussion). *Statistical Science 15*, 196–223.

Li, K.-C. (1992). On principal Hessian directions for data visualization and dimension reduction: Another application of Stein’s Lemma. *Journal of the American Statistical Association 87*, 1025–1040.

Porzio, G. C. and S. Weisberg (1999). Tests for lack-of-fit of regression models. Technical Report 634, School of Statistics, University of Minnesota.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics 12*, 1151–1172.