A Graphical Method for Assessing the Fit of a Logistic Regression Model

Iain Pardoe [*]
Charles H. Lundquist College of Business
University of Oregon

R. Dennis Cook [*]
School of Statistics
University of Minnesota

April 11, 2002

Abstract

Before a logistic regression model is used to describe the relationship between a binary response variable and predictors, the fit of the model should be assessed. The nature of any model deficiency may indicate that some aspect of the model should be reformulated or that poorly fitting observations need to be considered separately. We propose graphical methodology based on a Bayesian framework to address issues such as this. Publicly available software allows diagnostic plots to be constructed quickly and easily for any model of interest. These plots are more intuitive and meaningful than traditional graphical diagnostics such as residual plots.

Key Words: Bayesian methodology; Diagnostic plot; Marginal model plot; Model criticism; Posterior predictive distribution.

[*] Iain Pardoe is Assistant Professor, Department of Decision Sciences, Charles H. Lundquist College of Business, University of Oregon, Eugene, OR 97403–1208 (email: ipardoe@lcbmail.uoregon.edu); R. Dennis Cook is Professor, School of Statistics, University of Minnesota, Minneapolis, MN 55455 (email: dennis@stat.umn.edu). The authors thank the editor and two anonymous reviewers for helpful comments on an earlier version of this article. This research was partially supported by National Science Foundation grants DMS-9703777 and DMS-0103983.

1 INTRODUCTION

Cook and Pardoe (2000) suggested a graphical technique for assessing the fit of a regression model, which they called a "Gibbs marginal model plot." This methodology was developed for normal linear models and additive models in Pardoe (2001b), in which the plots were renamed "Bayes marginal model plots" (BMMPs). The plots provide a way for visualizing model uncertainty in the "marginal model plots" (MMPs) of Cook and Weisberg (1997).

This article describes the BMMP methodology in the context of a binary logistic regression analysis. Section 2 highlights the issues involved in assessing the fit of logistic regression models. It also introduces an example dataset on breast cancer diagnosis that will be used to illustrate the proposed methodology, outlines the difficulties of using residual plots, and reviews how MMPs can help with model assessment. However, without guidance on the level of uncertainty in the model, MMPs can be difficult to interpret. Section 3 provides details on how Bayesian model checking ideas can be used to address this problem, and illustrates how BMMPs can guide model improvement for the breast cancer data. We present a second example in Section 4, which follows an analysis by Bedrick, Christensen, and Johnson (1997) concerned with predicting survival at a trauma center. Section 5 contains a discussion.

2 BACKGROUND

2.1 Assessing "goodness of fit"

How can we assess the fit of a regression model that is to be used to explain the dependence of a binary response y on a vector of predictors $\boldsymbol{x}$, or to predict y from $\boldsymbol{x}$? Let the unknown conditional distribution of y given $\boldsymbol{x}$ be represented by its cumulative distribution function, $\mathrm{F}(y|\boldsymbol{x})$. Suppose we have a model for $\mathrm{F}(y|\boldsymbol{x})$, denoted by its cumulative distribution function, $\mathrm{M}_{\boldsymbol{\theta}}(y|\boldsymbol{x})$, where $\boldsymbol{\theta}$ is a vector of unknown parameters. Further, suppose we have completed the exploratory stage of the analysis, and have a reasonable first model that we would like to assess.

For a frequentist analysis, assume that $\boldsymbol{\theta}$ can be consistently estimated under $\mathrm{M}_{\boldsymbol{\theta}}(y|\boldsymbol{x})$ with $\widehat{\boldsymbol{\theta}}$. For a Bayesian analysis, assume that inference will be based on a posterior distribution for the model denoted by $\mathrm{M}_{\boldsymbol{\theta}}(y|\boldsymbol{x}, \boldsymbol{y}_d)$, where $\boldsymbol{y}_d$ is the n-vector of observed responses. Before using $\mathrm{M}_{\widehat{\boldsymbol{\theta}}}(y|\boldsymbol{x})$ or $\mathrm{M}_{\boldsymbol{\theta}}(y|\boldsymbol{x}, \boldsymbol{y}_d)$ to address a practical issue, we need to be confident that the model provides a *sufficiently accurate* approximation to $\mathrm{F}(y|\boldsymbol{x})$, where the accuracy is gauged relative to the practical issue. If the model is found to be deficient, the nature of the deficiency may indicate a need for some aspect of the model to be reformulated, or that poorly fitting or influential observations need to be considered separately.

Identifying the nature of a model deficiency for logistic regression is not an easy task. Pregibon (1981) developed the theory behind extensions of traditional linear regression diagnostics to logistic regression, and standard text-books, such as Agresti (2002) and Hosmer and Lemeshow (2000), offer some useful advice on model diagnostics in this context. However, the methodology on offer can be difficult to use, and the proposed graphical displays often lack clear, unambiguous meaning. Noting the difficulty of interpreting linear regression diagnostic displays in a logistic regression setting, Landwehr, Pregibon, and Shoemaker (1984) suggested modifications that led to pioneering work in this field. Eno and Terrell (1999) also proposed novel graphical techniques for logistic regression. We present an alternative graphical approach in Section 3.

2.2 Breast cancer data

The "Wisconsin Breast Cancer Data" (Bennett and Mangasarian, 1992) provides an example of a binary logistic model assessment problem. These data consist of 681 cases of potentially cancerous tumors, 238 of which turned out to be malignant, and 443 of which were benign. Determining whether a tumor is malignant or benign is traditionally accomplished with an invasive surgical biopsy procedure. An alternative, less invasive technique, allowing examination of a small amount of tissue from the tumor, is "Fine Needle Aspiration" (FNA). For the Wisconsin data, FNA provided nine cell features for each case; a biopsy was used to determine the tumor status as malignant or benign.

Features of the tissue cells can be used as predictors in a model with tumor status as the response. The hope is to use the model to successfully predict tumor status based only on the FNA predictors. Of critical importance is whether the model can provide an accurate alternative to the biopsy procedure for future patients.

The dataset consists of the following response and predictor variables:

y	= Class1	= 0 if malignant, 1 if benign	x_5	= Mitos	= mitoses
x_1	= Adhes	= marginal adhesion	x_6	= NNucl	= normal nucleoli
x_2	= BNucl	= bare nuclei	x_7	= Thick	= clump thickness
x_3	= Chrom	= bland chromatin	x_8	= UShap	= cell shape uniformity
x_4	= Epith	= epithelial cell size	x_9	= USize	= cell size uniformity

The predictors, $\boldsymbol{x} = (x_1, \ldots, x_9)^T$, are all integer values between one and ten (one represents a "normal" state, ten indicates a "most abnormal" state), and are determined by a doctor assessing the tissue cells through a microscope. Together, the predictors provide a wealth of information on tumor status. In fact, it appears that a subset of the predictors can provide nearly all the information available. Subset selection on the full set of nine predictors, removing the least significant predictor at each stage, leads to the following model worthy of consideration:

Model 1:
logit(Class1) = 11.049 -0.436 Adhes -0.470 BNucl -0.623 Chrom -0.378 NNucl -0.818 Thick

Some traditional numerical measures of fit include Wald p-values for predictors in the model each less than 0.0005, p-values for adding one more predictor each greater than 0.05, and residual deviance of 96.5 on 675 degrees of freedom. Based on these numbers, the model appears to fit well. However, perhaps plots of the data can give us further information on the fit of this model.

2.3 Residual plots

Two-dimensional plots of residuals versus fitted values or predictors are traditionally used to assess lack of fit of a regression model. The general idea is that if the model is correct then the sample residuals should appear independent of the predictors, with allowance for typically negligible dependence caused by substituting estimates for parameters. Consequently, observed patterns that indicate clear dependence also indicate violation of assumptions in the model. This paradigm

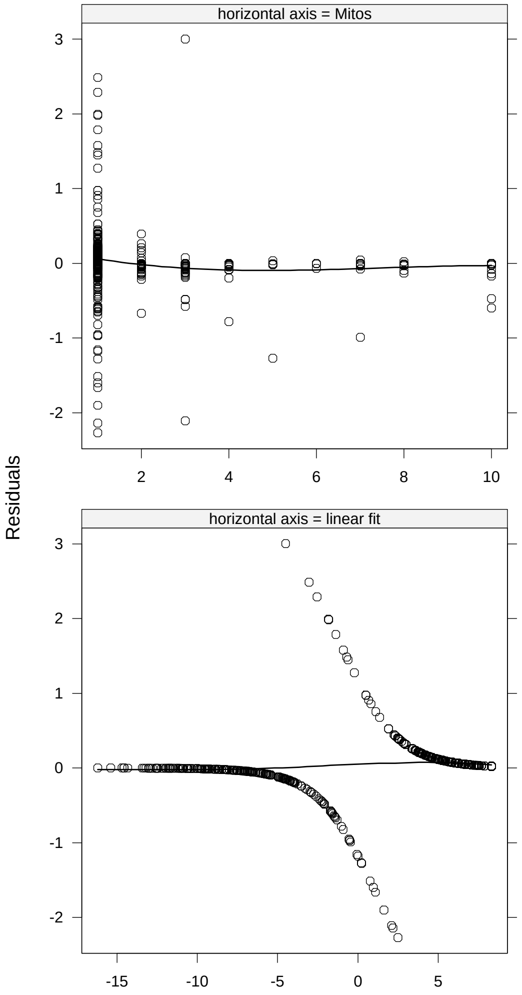

Figure 1: Residual plots for model 1 fit to the breast cancer data.

works best for linear models with additive errors. Complications arise in logistic regression and other generalized linear models because in such settings population residuals may not be independent of the predictors when the model is in fact correct.

Figure 1 shows two plots of deviance residuals with non-parametric smooths superimposed. These plots, and all subsequent plots, have been constructed using “Trellis Graphics” (Becker and Cleveland, 1996) in S-PLUS. For consistency with later plots, horizontal axis labels are shown in a strip at the top of the plots, and a common vertical axis label is displayed just once for all related plots. The upper plot is a residual plot with horizontal axis equal to Mitos, a predictor *not* in the model. If there are any unexpected patterns in this plot, then perhaps Mitos could usefully be added to the model. One problem with using residual plots in generalized linear models is that it can be difficult figuring out what kinds of patterns are unexpected and which are entirely to be expected. In binary logistic regression, the main unexpected pattern to look for is a non-constant mean function in the plot, as discussed by Cook (1998, sec. 15.1). Because the smooth of the residuals is flat relative to the variation in the residuals, there is apparently nothing to suggest that Mitos could usefully be included. However, it is less clear if the obvious pattern in the residuals contains relevant information or if a comparison of the variation in the smooth relative to variation in the residuals is in fact most appropriate.

The lower plot has horizontal axis equal to the linear fit from the model. This looks a little strange, and can be hard to interpret. In particular, the way the residuals fall on two distinct curves is an *expected* pattern due entirely to the fact that the response values are either zero or one, and the fitted probabilities are a non-linear but monotone function of the linear fit. But again the smooth of the residuals is essentially flat.

In both plots of Figure 1, we visually judged variation in the smooths relative to variation in the data using intuition from residual plots for linear, additive-error regressions. Such intuition may not be transferable to logistic regression. In addition, it is not always clear how residuals should be defined, both in a logistic regression context and from a Bayesian perspective (see Chaloner and Brant, 1988). Interpreting residual plots can be difficult because not all systematic patterns indicate a model deficiency, for example the two-curve pattern in the lower plot of Figure 1. Even when a systematic pattern does signal a problem, traditional interpretations can be misleading. Cook and Weisberg (1999) gave a continuous response regression example in which a fan-shaped plot of residuals versus fitted values results from an incorrectly specified mean function, not from non-constant variance as would usually be assumed. Such care with interpretation seems especially warranted with logistic regression, where there are several additional complications as discussed previously.

Residual plots are often judged to be useful in model assessment because systematic effects are removed from the data, apparently allowing any remaining structure to be detected more easily. However, constructing a residual can actually add structure and noise to the data, making it more difficult to discern model inadequacy. For example, consider an additive-error regression with true model $y = g(\boldsymbol{x}^T\boldsymbol{\theta}) + \epsilon$ where ϵ is independent of $\boldsymbol{x}$ and g is a nonlinear function. Suppose the OLS estimates of b_0 and $\boldsymbol{b}$ in the incorrect model $y = b_0 + \boldsymbol{x}^T\boldsymbol{b} + \epsilon$ converge to β_0 and $\boldsymbol{\beta}$. Then the population residuals are $r = g(\boldsymbol{x}^T\boldsymbol{\theta}) - \beta_0 - \boldsymbol{x}^T\boldsymbol{\beta} + \epsilon$. In general, $\boldsymbol{\beta} \not\propto \boldsymbol{\theta}$ so the residuals can depend on two linear combinations of the predictors, while y depends on only one linear combination. Detecting the model deficiency using residual plots may then be more difficult than using the original data because the presence of g can be obscured by $\boldsymbol{x}^T\boldsymbol{\beta}$. For instance, a plot of y versus $\boldsymbol{x}^T\boldsymbol{\theta}$ will have mean function $g(\boldsymbol{x}^T\boldsymbol{\theta})$ with constant variance, $\mathrm{Var}(y|\boldsymbol{x}^T\boldsymbol{\theta}) = \mathrm{Var}(\epsilon)$. But a plot of r versus $\boldsymbol{x}^T\boldsymbol{\theta}$ will have mean function $g(\boldsymbol{x}^T\boldsymbol{\theta}) - \beta_0 - \mathrm{E}(\boldsymbol{x}^T\boldsymbol{\beta}|\boldsymbol{x}^T\boldsymbol{\theta})$ and variance function

$$\mathrm{Var}(r|\boldsymbol{x}^T\boldsymbol{\theta}) = \mathrm{Var}(\boldsymbol{x}^T\boldsymbol{\beta}|\boldsymbol{x}^T\boldsymbol{\theta}) + \mathrm{Var}(\epsilon) \geq \mathrm{Var}(y|\boldsymbol{x}^T\boldsymbol{\theta})$$

Residual plots work best when $\boldsymbol{\beta} \propto \boldsymbol{\theta}$ so that $\mathrm{Var}(\boldsymbol{x}^T\boldsymbol{\beta}|\boldsymbol{x}^T\boldsymbol{\theta}) = 0$ and consequently $\mathrm{Var}(r|\boldsymbol{x}^T\boldsymbol{\theta}) = \mathrm{Var}(y|\boldsymbol{x}^T\boldsymbol{\theta})$. Nevertheless, when $\mathrm{Var}(r|\boldsymbol{x}^T\boldsymbol{\theta}) > \mathrm{Var}(y|\boldsymbol{x}^T\boldsymbol{\theta})$, dependence on $\boldsymbol{x}^T\boldsymbol{\theta}$ will typically be less clear in a plot of r versus $\boldsymbol{x}^T\boldsymbol{\theta}$ than in a plot of y versus $\boldsymbol{x}^T\boldsymbol{\theta}$.

The same type of situation occurs in logistic regression, but the effects are more complicated because of the non-additive nature of the errors. Further discussion of the issues involved with interpreting residual plots is available in Cook (1994), Cook and Weisberg (1997), and Cook and Weisberg (1999).

2.4 Marginal model plots

Alternatively, we can visualize *goodness* of fit in a marginal model plot (MMP) such as that shown in Figure 2 with horizontal axis $h =$ Mitos. Cook and Weisberg (1997) introduced these plots from

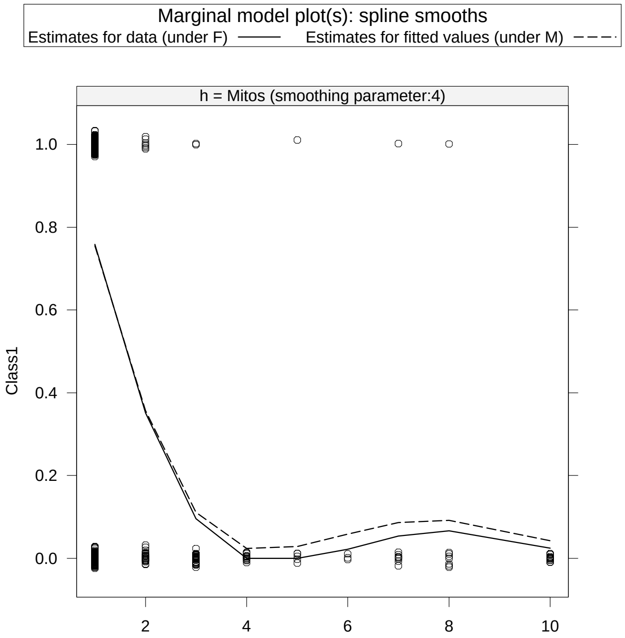

Figure 2: MMP for the mean with $h =$ Mitos for model 1 fit to the breast cancer data.

a frequentist perspective, in which the solid line is a smooth of the data and the dashed line is a smooth of the fitted values from the model. The rationale for the MMP is this statement:

$$\mathrm{E_F}(y|\boldsymbol{x}) = \mathrm{E}_{\widehat{\mathrm{M}}}(y|\boldsymbol{x}), \quad \forall\, \boldsymbol{x} \in \mathcal{X} \subset \mathrm{I\!R}^p \tag{1}$$

$$\iff \mathrm{E_F}(y|h) = \mathrm{E}_{\widehat{\mathrm{M}}}(y|h), \quad \forall\, h = h(\boldsymbol{x}) : \mathrm{I\!R}^p \to \mathrm{I\!R}^{\mathit{1}} \tag{2}$$

where $\mathrm{E_F}$ denotes expectation under F, $\mathrm{E}_{\widehat{\mathrm{M}}}$ denotes expectation under $\mathrm{M}_{\widehat{\boldsymbol{\theta}}}$, and $\mathcal{X}$ is the sample space of $\boldsymbol{x}$. Think of $\boldsymbol{x}$ here in the same way that it is thought about in subset selection, i.e. predictors that are included in the model being considered, as well as potential predictors not in the current model. This result, which follows from Proposition 4.3 in Cook (1998), requires only that h be measurable with respect to $\boldsymbol{x}$. Residual plots also rely on (1) and (2) for their interpretation, except y is replaced with the residual.

Equality (1) is what we would like to check, but if the dimension of $\boldsymbol{x}$ is greater than two, then $\mathrm{E}(y|\boldsymbol{x})$ can be difficult to visualize. However, because h is univariate, $\mathrm{E}(y|h)$ can be visualized in a 2-D scatterplot, and equality (2) can be checked. So, the idea in a MMP is to compare $\mathrm{E_F}(y|h)$ and $\mathrm{E}_{\widehat{\mathrm{M}}}(y|h)$ for various h to gain information about the relationship between $\mathrm{E_F}(y|\boldsymbol{x})$ and $\mathrm{E}_{\widehat{\mathrm{M}}}(y|\boldsymbol{x})$. The mean function based on F can be thought of as *model-free*, while that based on $\widehat{\mathrm{M}}$ can be thought of as *model-based*.

We can estimate the two mean functions with smooths. Obtain $\widehat{\mathrm{E}}_{\mathrm{F}}(y|h)$ by smoothing y versus h using a non-parametric smooth such as a cubic smoothing spline. The corresponding model-based estimate of the mean function uses the relationship $\mathrm{E}_{\widehat{\mathrm{M}}}(y|h) = \mathrm{E}[\mathrm{E}_{\widehat{\mathrm{M}}}(y|\boldsymbol{x})|h]$. So, obtain $\widehat{\mathrm{E}}_{\widehat{\mathrm{M}}}(y|h)$ by smoothing $\mathrm{E}_{\widehat{\mathrm{M}}}(y|\boldsymbol{x})$ versus h; note that $\mathrm{E}_{\widehat{\mathrm{M}}}(y|\boldsymbol{x})$ is the (assumed) mean function from the fitted model, or—in other words—the *fitted values* from the model. Superimpose $\widehat{\mathrm{E}}_{\mathrm{F}}(y|h)$ and $\widehat{\mathrm{E}}_{\widehat{\mathrm{M}}}(y|h)$ on a plot of y versus h to obtain a MMP for the mean in the (marginal) direction h. Using the same method and smoothing parameter for the mean function estimates under F and $\widehat{\mathrm{M}}$

allows their point-wise comparison, since any estimation bias should approximately cancel. Even though one mean function estimate smooths (binary) data and the other smooths (continuous) fitted values, each smooth estimates the probability that y is one as a function of h, and each smooth has approximately the same bias. See Bowman and Young (1996) for elaboration of this point.

Ideas for selecting useful functions h to consider in practice are given in Cook and Weisberg (1997), and include fitted values, individual predictors in the model, potential predictors not in the model, linear combinations of the predictors, and random linear projections of the predictors. Other possibilities include functions h where lack of fit is most likely to be observed. Promising candidates include functions found using *sliced inverse regression* (Li, 1991), *principal Hessian directions* (Li, 1992), and *sliced average variance estimation* (Cook and Weisberg, 1991).

Now, if M is an accurate approximation to F, then for any h the marginal mean function 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 mean functions do not agree for one particular h calls M into question; if they agree for a variety of plots, there is support for M.

So, how should Figure 2 be interpreted? In this plot, the smoothing splines have four effective degrees of freedom (the smoothing parameter for smoothing splines) and the points have been jittered to aid visualization of data density. Most of the data is on the left where Mitos is equal to one or two, and here the smooths match well. But, for Mitos three or higher, the model seems to predict higher probabilities of a tumor being benign than the data indicate. But, is the gap between the smooths so large that we should be concerned, or so small that we can just put it down to random variation?

The same issue of variability arises in residual plots also. For instance, the smooths in Figure 1 are judged against horizontal lines at 0. The deviance, which is essentially a numerical summary of a residual plot for a logistic regression, provides one way to address this issue. However, if a model is identified as poorly fitting due to a high deviance in relation to the error degrees of freedom, there is no guidance available on how to improve the model. It would be helpful to *see* the nature of the lack-of-fit in a graphical display, and in this respect the MMP is to be preferred over the residual plot since it is easier to interpret and therefore potentially more informative.

Even if $\mathrm{M}_{\boldsymbol{\theta}}(y|\boldsymbol{x}) = \mathrm{F}(y|\boldsymbol{x})$, the estimated marginal mean function estimates in an MMP would not match exactly. From a frequentist perspective, the data can be thought of as just one realization of many possible samples. So, a possible solution to the problem of comparing the estimates is to calculate a sampling-theory confidence band or perhaps generate replicate data by bootstrapping. Alternatively, from a Bayesian perspective, the data are fixed, but the variability in the model estimates is given explicitly by the posterior distribution for the parameters. A possible solution to the assessment problem displays this variability in the model smooth, allowing the analyst to more easily judge whether it would be reasonable for the data to be generated by the model in question.

3 BAYES MARGINAL MODEL PLOTS

3.1 Visualizing model uncertainty

To introduce ideas and keep notation concise, consider assessing how well a model $\mathrm{M} = \mathrm{M}(y|\boldsymbol{\theta})$ fits 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 M based on the marginal, or predictive, distribution of $\boldsymbol{y}$. He suggested assessing M by referring the value of the predictive density for the observed data, $f(\boldsymbol{y}_d|\mathrm{M})$, to the density function $f(\boldsymbol{y}|\mathrm{M})$, by calculating a tail area, say. A "small" tail area indicates that $\boldsymbol{y}_d$ would be unlikely to have been generated by M, and thus calls M into question. More generally, M can be assessed by referring the value of the predictive density of some relevant checking function, $g(\boldsymbol{y})$, at $\boldsymbol{y}_d$ to its predictive density, for a variety of g. Examples of useful g in practice include residuals, order statistics, and moment estimators.

Rubin (1984) proposed an alternative approach that does not require proper priors, as Box's approach does, using the posterior predictive density

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

where $f(\boldsymbol{y}|\boldsymbol{\theta}, \mathrm{M})$ is the likelihood for $\boldsymbol{y}$ and $\pi(\boldsymbol{\theta}|\boldsymbol{y}_d, \mathrm{M})$ is the posterior density of $\boldsymbol{\theta}$. The posterior predictive distribution of $\boldsymbol{y}$ can be thought of as a distribution for potential data that we might observe, if the model that we think produced $\boldsymbol{y}_d$, *including the particular $\boldsymbol{\theta}$ value*, was used to produce a new set of data. Since this particular $\boldsymbol{\theta}$ value is unknown, average over plausible values using its posterior distribution. Again, diagnostics similar to Box's tail area and checking functions g can be constructed. Use of the posterior predictive distribution in a goodness of fit test was first proposed by Guttman (1967). Rubin's approach has been extended by Meng (1994) to allow the checking function g to depend on nuisance parameters as well as on $\boldsymbol{y}$, and by Gelman, Meng, and Stern (1996) to allow g to also depend on $\boldsymbol{\theta}$.

Another way to think about Rubin's approach is in terms of a sampling simulation. Gelman et al. (1996) provide references to many papers that discuss this interpretation. The idea is to draw a value of $\boldsymbol{\theta}$ from its posterior distribution, and then generate a sample of n realizations from the model 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 M. Then, intuitively, if $\boldsymbol{y}_d$ "looks like" a typical realization from M, there is no reason to doubt the fit of M. On the other hand, if $\boldsymbol{y}_d$ appears to be very "unusual" with respect to the m realizations from M, then M is called into question. To do this in practice, methods for comparing $\boldsymbol{y}_d$ to the m realizations from 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 posterior distribution for $\boldsymbol{\theta}$.

A graphical way to do this is based on the MMPs introduced earlier. In regression, $\boldsymbol{\theta}$ provides "fitted values". So, instead of sampling y, compare model-free predicted values with expected y-values based on sampled $\boldsymbol{\theta}$ values. A Bayes marginal model plot (BMMP) is a scatterplot of y versus h with a mean function estimate under F superimposed. Then, instead of also superimposing the mean function estimate under $\widehat{\mathrm{M}}$ on this plot, superimpose a mean function estimate for each model sample $\mathrm{M}_{\boldsymbol{\theta}_t}$, $t = 1, \ldots, m$.

Recall that the smoothing parameters for the smooths in a particular MMP need to be equal to allow their point-wise comparison. Similarly, the smooths in a BMMP should all have the same smoothing parameter, γ. Therefore it is desirable to select γ so that the smooths are flexible enough to capture clear systematic trends in all the corresponding scatterplots, while not over-fitting too much in any one scatterplot, over-reacting to individual points, and tracking spurious patterns. This is clearly impractical, but a viable alternative is to graphically select γ to capture the systematic trends in both a scatterplot of the data (y) versus h and a scatterplot of the fitted values from the model versus h. Further discussion of this issue is given in Section 5.

If enough samples are taken, say $m = 100$, the Bayes mean function estimates, $\widehat{\mathrm{E}}_{\mathrm{M}_{\boldsymbol{\theta}_t}}(y|h)$, $t = 1, \ldots, m$, will form a mean function *band* under M. The plot then provides a visual way of determining whether there is any evidence to contradict the possibility that $\mathrm{F}(y|\boldsymbol{x}) = \mathrm{M}(y|\boldsymbol{x})$. If, for a particular h, the mean function estimate under F lies *substantially outside* the mean function band under M *or* it does not follow the general pattern shown by the model smooths, then M is called into question. If, no matter what the function h is, the mean function estimate under F lies *broadly inside* the mean function band under M *and* it follows the general pattern shown by the model smooths, then perhaps M provides an accurate description of the conditional distribution of $y|\boldsymbol{x}$ and is a useful model.

The binary logistic regression model can be written

$$\begin{aligned} y_i|(\boldsymbol{x}_i, p_i) &\sim \text{Bernoulli}(p_i) \\ p_i &= \Pr(y = 1|\boldsymbol{x}_i) = \mathrm{E}(y|\boldsymbol{x}_i) \\ \text{logit}(p_i) &= \log\left(\frac{p_i}{1 - p_i}\right) = \boldsymbol{\theta}^T\boldsymbol{x}_i \end{aligned}$$

One possible prior for this regression is

$$\boldsymbol{\theta} \sim \mathrm{N}(\mathbf{0}_p, k\boldsymbol{I}_p) \tag{3}$$

where k can be set to reflect the degree of prior uncertainty for any particular dataset. It is not possible to sample directly from the posterior, so instead Markov chain simulation can be used to obtain the samples. In particular, posterior samples can be obtained by Gibbs sampling using WinBUGS software (Spiegelhalter, Thomas, and Best, 1999). Checking convergence in Markov chain sampling is very important, and software is available from various sources to assist in this task. Some software that works well with WinBUGS output is BOA (Smith, 2001).

Constructing a BMMP for the mean in direction h requires model-free and model-based estimates of the mean function with respect to h. To obtain the model-free estimate $\widehat{\mathrm{E}}_{\mathrm{F}}(y|h)$, smooth the data $\{y_i\}$ on $\{h_i\}$. To obtain the model-based estimates $\widehat{\mathrm{E}}_{\mathrm{M}_{\boldsymbol{\theta}_t}}(y|h)$, smooth the fitted-values based on the posterior samples $\{\mathrm{E}_{\mathrm{M}_{\boldsymbol{\theta}_t}}(y|\boldsymbol{x}_i)\}$ on $\{h_i\}$. The fitted values corresponding to posterior samples $\boldsymbol{\theta}_t$ are

$$\mathrm{E}_{\mathrm{M}_{\boldsymbol{\theta}_t}}(y|\boldsymbol{x}_i) = p_{it} = \frac{1}{1 + \exp(-\boldsymbol{\theta}_t^T\boldsymbol{x}_i)}, \quad i = 1, \ldots, n;\ t = 1, \ldots, m$$

3.2 Breast cancer data revisited

The BMMP equivalent to Figure 2 is shown in Figure 3. In this plot, the prior uncertainty parameter, k, in (3) was set to be 10^6, the smoothing splines have four effective degrees of freedom, and $m = 100$. Here, the black smooth of the data lies below the gray band of the fitted probability smooths for values of Mitos three or higher. Mitos clearly adds information on the probability of being benign not provided by the five predictors in the model. This plot, in contrast to a residual plot, can be interpreted straightforwardly, incorporates model uncertainty, and provides guidance on model improvement. Recall that the Wald p-values for adding one more predictor to the model were each greater than 0.05; the BMMP tells us that we should not be so hasty in neglecting Mitos because of this. So, let's add Mitos to the model to see what happens.

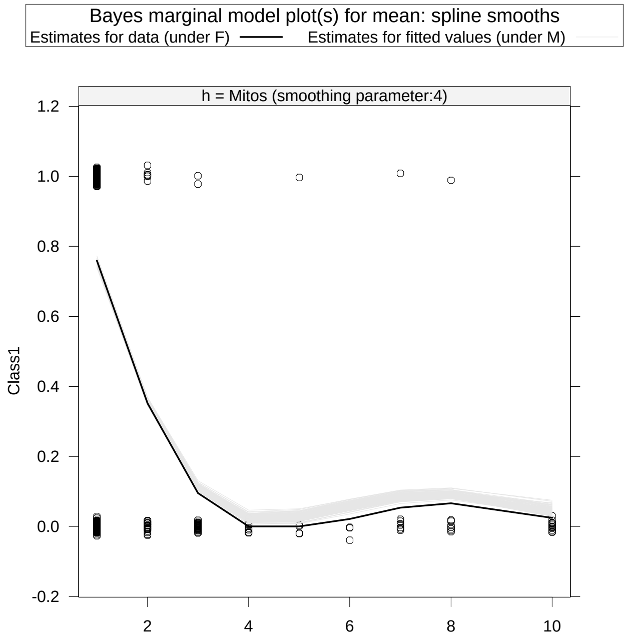

Figure 3: BMMP for the mean with $h =$ Mitos for model 1 fit to the breast cancer data.

Model 2:
logit(Class1) = 11.473 $-$0.445 Adhes $-$0.474 BNucl $-$0.633 Chrom $-$0.362 NNucl $-$0.722 Thick $-$0.682 Mitos

For this model, the BMMP for the mean with $h =$ Mitos is the upper plot of Figure 4. In this plot, the smoothing splines again have four effective degrees of freedom. This plot shows a big improvement over Figure 3, so it appears that adding Mitos to the model is useful. However, recall that a *series* of BMMPs needs to be considered in order to be confident in the model. So, how about the BMMP for the mean with h set equal to the linear fit from the model? This is the lower plot of Figure 4. In this plot, the smoothing splines have twelve effective degrees of freedom—increased flexibility in the smooths is needed to fit the "logistic curve" shapes of the data and the fitted probabilities. The black smooth of the data lies mostly inside the gray band of the fitted probability smooths, but it gets very close to the edge of the band at one point. The model appears to fit most of the data very well, but has trouble with cases "in the middle" when the linear fit is close to zero. Reactions to this behavior in the middle are context dependent, and Section 5 contains some discussion of a numerical summary measure for the plot that could assist in cases such as this. Nonetheless, BMMPs have taken us much further in model assessment and understanding for this dataset than either residual plots or MMPs.

4 TRAUMA DATA

Bedrick et al. (1997) gave a detailed Bayesian analysis of a model for predicting survival at a trauma center. They analyzed data on 278 survivors (LIVE $= 1$) and 22 fatalities (LIVE $= 0$) with the following predictor variables:

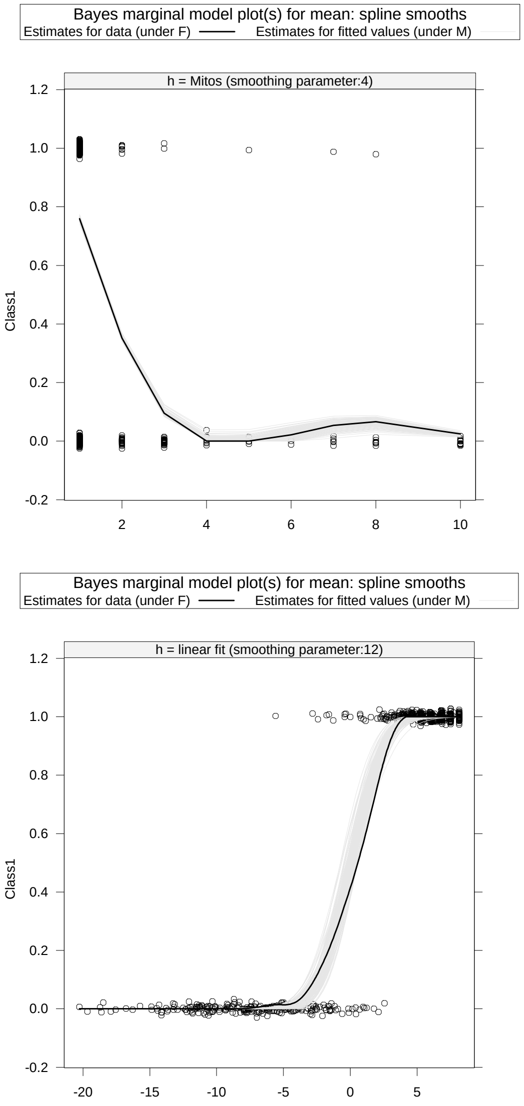

Figure 4: BMMP for the mean for model 2 fit to the breast cancer data: $h =$ Mitos (upper) and $h =$ the linear fit (lower).

ISS	= injury severity score	= 0 if no injuries, up to 75 if severely injured
RTS	= revised trauma score	= 0 if no vital signs, up to 7.84 if normal vital signs
AGE	= age in years	= from 1 to 94
TI	= type of injury	= 0 if blunt, 1 if penetrating
AGE.TI	= AGE × TI interaction	= from 0 to 94

Bedrick et al. provide a very nice discussion of modeling these data, including elicitation of expert prior opinion, inference for survival probability, and some model-checking diagnostics. The discussion related to this latter point is restricted to case deletion diagnostics however, and the reader is perhaps left with a vague uneasiness about whether this inventive model does indeed provide a sufficiently accurate approximation to the conditional distribution of LIVE given the available predictors.

We fitted a logistic model essentially identical to that of Bedrick et al. using WinBUGS soft-

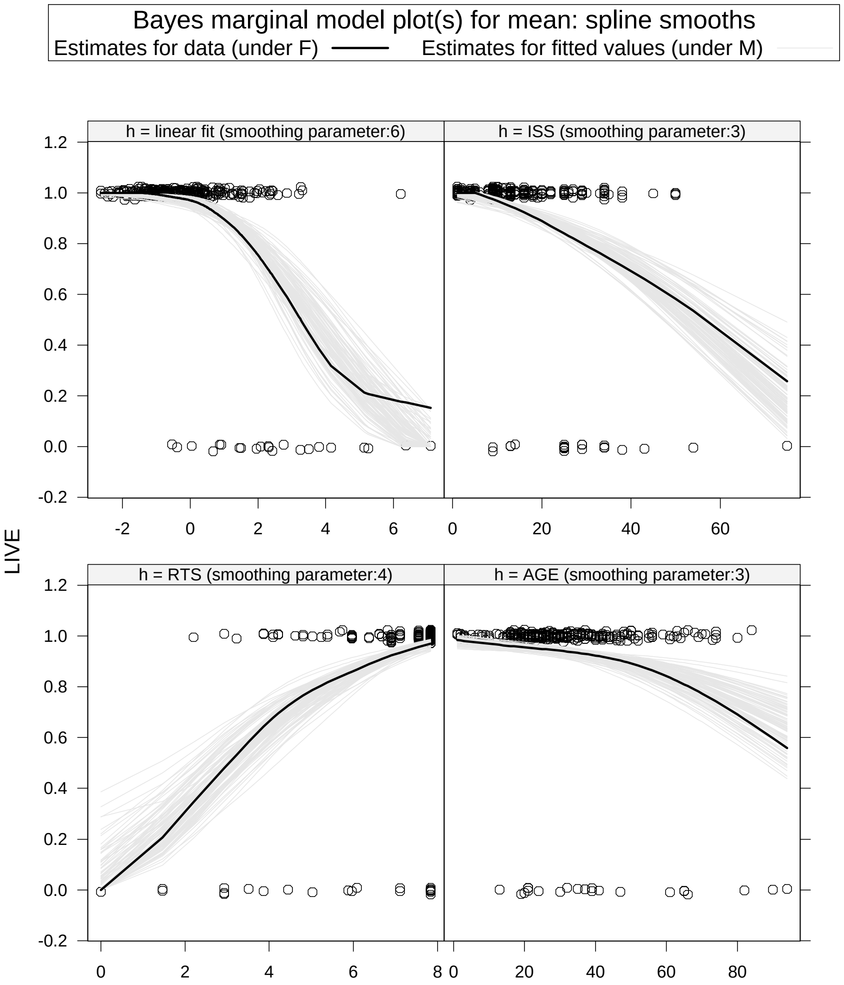

Figure 5: BMMPs for the mean for the trauma data.

ware. Our priors were restricted to integer-parameter beta distributions, whereas the original analysis utilized non-integer values for the prior beta distributions. Nevertheless, posterior distribution summaries were in close agreement. The model gives BMMPs for the mean with h set to four different functions as shown in Figure 5. In these plots, the smoothing splines have six, three, four, and three effective degrees of freedom for the linear fit, ISS, RTS, and AGE. The black model-free smooths go mostly through the middle of the gray model-based smooths for each of these h-functions. Other h-functions (not shown) also display this feature. This model does indeed appear to fit very well.

The BMMP for the linear fit indicates some lack of fit on the far right of the plot. This lack of fit is being driven by the case in the top right corner. This is case 232 who survived despite all indications to the contrary. Bedrick et al. identified this case through case deletion diagnostics. Figure 5 enables an interpretation of this case's influence on the model in relation to predicted probability of survival.

5 DISCUSSION

BMMPs offer a quick and easy way to check models graphically. The sampling needs to be done only once for each model and cycling through BMMPs in a variety of directions h provides guidance on the fit of the model. In an area where we are unaware of any other competing graphical methods, we propose this methodology as a viable alternative to the residual plot that avoids many of the latter's drawbacks in relation to definition and interpretation. In particular, BMMPs can be used to aid model assessment (as opposed to model selection—BMMPs are not designed for subset selection, for example).

5.1 Bayes discrepancy measure

A discrepancy measure could perhaps provide a useful numerical complement to a BMMP to aid its interpretation. Consider a discrepancy measure for a BMMP for the mean in direction h based on the average squared distance between the model smooths and the data smooth

$$\mathrm{D_M}(h) = \mathrm{E}_{\boldsymbol{\theta}}[\mathrm{E}_h\{(\mathrm{E_F}(y|h) - \mathrm{E}_{\mathrm{M}_{\boldsymbol{\theta}}}(y|h))^2\}]$$

where $\mathrm{E}_{\boldsymbol{\theta}}$ represents expectation with respect to the posterior distribution of $\boldsymbol{\theta}$. $\mathrm{D_M}(h)$ can be estimated using the posterior $\boldsymbol{\theta}$ samples and the smooths $\widehat{\mathrm{E}}_\mathrm{F}(y|h)$ and $\widehat{\mathrm{E}}_{\mathrm{M}_{\boldsymbol{\theta}_t}}(y|h),\ t = 1, \dots, m$, by

$$\mathrm{D}_{\widehat{\mathrm{M}}_\mathrm{d}}(h) = \frac{1}{nm}\sum_{i=1}^{n}\sum_{t=1}^{m}\left(\widehat{\mathrm{E}}_\mathrm{F}(y|h_i) - \widehat{\mathrm{E}}_{\mathrm{M}_{\boldsymbol{\theta}_t}}(y|h_i)\right)^2$$

The "null" distribution of this discrepancy measure can be estimated empirically using just the model smooths

$$\mathrm{D}_{\widehat{\mathrm{M}}_\mathrm{j}}(h) = \frac{1}{n(m-1)}\sum_{i=1}^{n}\sum_{t=1}^{m}\left(\widehat{\mathrm{E}}_{\mathrm{M}_{\boldsymbol{\theta}_j}}(y|h_i) - \widehat{\mathrm{E}}_{\mathrm{M}_{\boldsymbol{\theta}_t}}(y|h_i)\right)^2 \quad j = 1, \dots, m$$

One of the denominators here is $m - 1$ rather than m since the distance between a model smooth and itself is identically zero. Then, the *Bayes discrepancy measure* is:

$$\text{BDM}(h) = \frac{1}{m}\sum_{j=1}^{m}\mathrm{I}\left(\mathrm{D}_{\widehat{\mathrm{M}}_\mathrm{j}}(h) > \mathrm{D}_{\widehat{\mathrm{M}}_\mathrm{d}}(h)\right)$$

This measure calculates the proportion of m model smooth discrepancies, $\mathrm{D}_{\widehat{\mathrm{M}}_\mathrm{j}}(h)$, that are larger than the data smooth discrepancy, $\mathrm{D}_{\widehat{\mathrm{M}}_\mathrm{d}}(h)$. Intuitively, it essentially counts the number of model smooths that are farther away (on average) from their companions than the data smooth is from them. The larger this measure, the more there is support for the model (in direction h), since the data smooth does not appear to be very "unusual" with respect to the model smooths. As this measure becomes smaller, there is less support for the model, since now the data smooth appears more unusual. Our experience using this methodology suggests that a BDM value less than 0.05 for any h strongly indicates lack of fit, while a BDM value between 0.05 and 0.10 indicates that improvement in the model may be possible.

For BMMPs to be useful graphical model assessment methodology, they should show no lack of fit for all h when the model is correct, but show clear lack of fit for at least one h when the model is incorrect. For BDMs to be a useful numerical complement to BMMPs, they should be far from zero for all h when the model is correct, but be close to zero for at least one h when the model is incorrect. In addition, choice of smoother type, smoothing parameter, number of samples, and plotting direction h, could each have an effect on the conclusions of the procedure. Details of simulations designed to consider these issues in a variety of regression situations were given by Pardoe (2001a). Some broad conclusions follow.

5.2 Smoothing issues

BMMPs utilize nonparametric scatterplot smoothers, and both cubic smoothing splines and loess smoothers performed well. Other smoother methods, including kernel smooths and Friedman's "super smoother," performed less well. All the preceding techniques make use of standard nonparametric methods for continuous data. These ignore the binary nature of the response variable in the current application. There are specialized smoothers for binary data, for example the local likelihood smoother in Bowman and Azzalini (1997, page 53), that could be used as an alternative. However, such smoothers are much more computer-intensive and at the present time impractical for use in BMMPs. Also, it is our experience that use of more sophisticated smoothers usually changes the visual impression of a BMMP very little, and so use of standard continuous data smoothers does not seem unreasonable.

Whichever smoothing method is used, some care must be exercised in the selection of the smoothing parameter. Over-smoothing can result in smooths that miss clear patterns, while under-smoothing can produce highly variable smooths that track spurious patterns. Broadly speaking, when the smoothing parameter was appropriately chosen as a reasonable compromise between over- and under-smoothing, the simulations correctly diagnosed poorly-fitting models and well-fitting models most of the time. However, it was sometimes possible to over-smooth an incorrect model and miss evidence of lack of fit, and to under-smooth a correct model and falsely conclude model inadequacy.

There appear to be no reliable short-cut methods for pre-selecting the smoothing parameter ahead of time, or recommended values that will work in most situations. The analyst needs to look at the scatterplots of the data (y) versus h and of the fitted values from the model versus h and choose a smoothing parameter that produces reasonable smooths in both plots simultaneously. If in doubt, err on the side of over-smoothing rather than under-smoothing, since under-smoothing sometimes produced low BDM values in cases when the model was good, but over-smoothing rarely lead to high BDM values in cases where the model was bad.

The amount of over- or under-smoothing has to be quite extreme to produce poor results however. For example, the effect of decreasing or increasing the effective degrees of freedom by one for the smoothing splines in Figure 3 and the lower left plot of Figure 5 can be seen in Figures 6 and 7. The qualitative nature of the plots is very similar to that of the earlier plots. The over-smoothed upper plot of Figure 6 continues to indicate lack of fit for values of Mitos equal to three, four or five; interpretation is more difficult for larger values of Mitos since the data smooth is effectively truncated at zero. The under-smoothed lower plot of Figure 6 conveys an identical visual impression to that of Figure 3 even though the smooths are more wiggly. The over-smoothed upper plot of Figure 7 continues to indicate good fit for $h =$ RTS, as does the under-smoothed lower-plot,

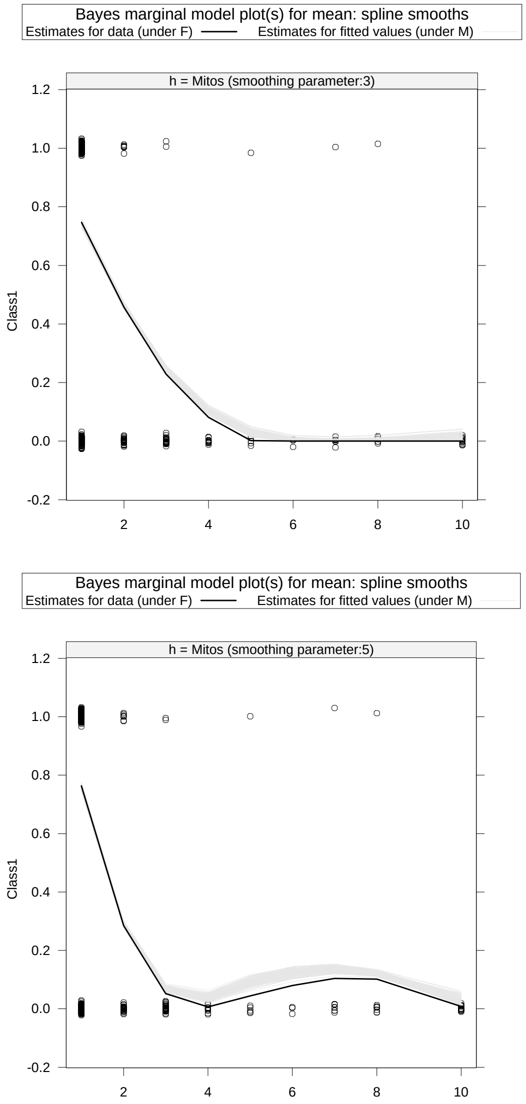

Figure 6: BMMPs for the mean with $h =$ Mitos for model 1 fit to the breast cancer data: smoothing parameter = 3 (upper) and smoothing parameter = 5 (lower).

although again the smooths are more wiggly.

The suggested method for choosing the smoothing parameter relies on human perception of the patterns in the plots of the data (y) versus h and of the fitted values from the model versus h. Our experience with the methodology indicates that, in the case of smoothing splines, variation of two or three effective degrees of freedom between different analysts is not unreasonable, but, as illustrated above, makes little (qualitative) difference to the plots. Varying the smoothing parameter any more than this can have adverse effects, but is unlikely to be a problem in practice since it would likely correspond to a poor representation of the patterns in the data and the fitted values.

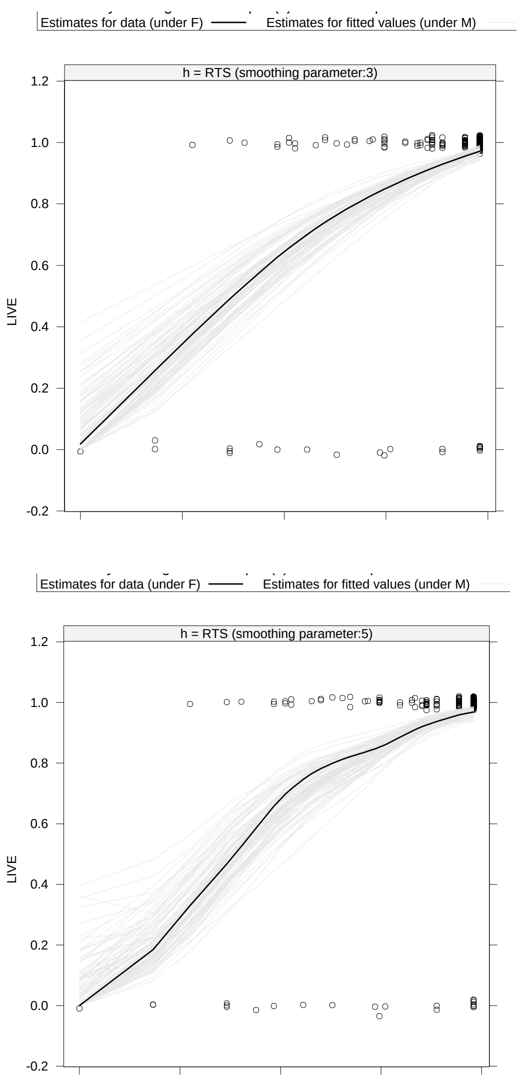

Figure 7: BMMPs for the mean with $h =$ RTS for model fit to the trauma data: smoothing parameter = 3 (upper) and smoothing parameter = 5 (lower).

5.3 Number of posterior samples

The actual number of samples used to create the BMMP does not appear to have a large impact on this methodology, so there would appear to be no need to use any more than 100 samples for each plot. Using substantially less than 100 samples would likely lead to poor resolution in the plots and make interpretation difficult.

5.4 Extensions

The examples considered here adopted Rubin's approach using posterior sampling. BMMPs based on Box's approach using prior sampling could be constructed similarly, although their interpretation would be a little different. An intermediate approach using cross-validation/jack-knifing ideas might also be useful, although implementation becomes much trickier computationally. One possible implementation is via re-weighting of a regular posterior sample (see Chib and Geweke, 2001, for example).

Details for other types of regression models, such as linear and additive models, follow from the discussion for the binary logistic model. Other models—for example survival models, time series models, and random effects models—could no doubt benefit from the application of the ideas in this paper. One strength of the BMMP methodology is that it is broadly applicable to *any* regression situation, with just the details of obtaining samples and constructing the actual plots to worry about.

In addition, 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); and 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. For example, as suggested by a referee, and also proposed in Pardoe (2001a, sec. 8.3), a “Bayesian residual plot” could consist of smooths of many simulated models’ residuals using a sample from the posterior distribution of model parameters. The sample of smooths either would or would not cover the constant function $\mathrm{E}(\text{residual}|h) = 0$. However, in interpreting such a plot, careful consideration of the issues raised in Section 2.3 would have to be given. Also, adding posterior-based smooths to the upper plot of Figure 1 seems unlikely to indicate the model deficiency that is apparent from the BMMP in Figure 3.

5.5 Software

S-PLUS and R functions have been developed that can be used in conjunction with WinBUGS and BOA to construct BMMPs for the mean in any user specified direction h. The software is available at `http://lcb1.uoregon.edu/ipardoe/research/bmmpsoft.htm` and further details are provided in Pardoe (2001c).

References

Agresti, A. (2002). *Categorical Data Analysis* (2nd ed.). New York: Wiley.

Becker, R. A. and W. S. Cleveland (1996). *S-PLUS Trellis Graphics User’s Manual*. Murray Hill, NJ: Bell Labs.

Bedrick, E. J., R. Christensen, and W. Johnson (1997). Bayesian binomial regression: Predicting survival at a trauma center. *The American Statistician 51*, 211–218.

Bennett, K. P. and O. L. Mangasarian (1992). Robust linear programming discrimination of two linearly inseparable sets. In *Optimization Methods and Software*, Volume 1, pp. 23–34. Gordon and Breach Science Publishers.

Bowman, A. W. and A. Azzalini (1997). *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-PLUS Illustrations*. Oxford: Oxford University Press.

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

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.

Chaloner, K. M. and R. Brant (1988). A Bayesian approach to outlier detection and residual analysis. *Biometrika 75*, 651–659.

Chib, S. and J. Geweke (2001). Bayesian Analysis, Computation and Communication software (BACC). Available at http://www.econ.umn.edu/∼bacc.

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

Cook, R. D. (1994). On the interpretation of regression plots. *Journal of the American Statistical Association 89*, 177–189.

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.

Cook, R. D. and S. Weisberg (1999). Graphs in statistical analyses: Is the medium the message? *The American Statistician 53*, 29–37.

Eno, D. R. and G. R. Terrell (1999). Scatterplots for logistic regression (with discussion). *Journal of Computational and Graphical Statistics 8*, 413–430.

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.

Guttman, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society, Series B (Methodological) 29*, 83–100.

Hosmer, D. W. and S. Lemeshow (2000). *Applied Logistic Regression* (2nd ed.). Wiley.

Landwehr, J. M., D. Pregibon, and A. C. Shoemaker (1984). Graphical methods for assessing logistic regression models (with discussion). *Journal of the American Statistical Association 79*, 61–83.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association 86*, 316–342.

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.

Meng, X.-L. (1994). Posterior predictive p-values. *The Annals of Statistics 22*, 1142–1160.

Pardoe, I. (2001a). *A Bayesian Approach to Regression Diagnostics*. Ph. D. thesis, School of Statistics, University of Minnesota.

Pardoe, I. (2001b). A Bayesian sampling approach to regression model checking. *Journal of Computational and Graphical Statistics 10*, 617–627.

Pardoe, I. (2001c). User's manual for `bmmp` S-PLUS and R software. Technical Report 639, School of Statistics, University of Minnesota.

Pregibon, D. (1981). Logistic regression diagnostics. *The Annals of Statistics 9*, 705–724.

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

Smith, B. J. (2001). Bayesian Output Analysis Program (BOA). Version 1.0.0 for S-PLUS and R, available at http://www.public-health.uiowa.edu/boa.

Spiegelhalter, D. J., A. Thomas, and N. G. Best (1999). *WinBUGS Version 1.2 User Manual*. Cambridge, UK: MRC Biostatistics Unit.