Sampling to Assess the Fit of Regression Models

Iain Pardoe and R. Dennis Cook
University of Minnesota

Graphical Regression Diagnostics

Consider a regression problem with n observations of a univariate response, $\boldsymbol{y} = (y_1, \ldots, y_n)^T$, and p predictors, $\boldsymbol{x} = (x_1, \ldots, x_p)^T$. Suppose a model $\mathrm{M} = \mathrm{M}(\boldsymbol{\theta})$ has been derived for the conditional cumulative distribution function F of y given $\boldsymbol{x}$, where $\boldsymbol{\theta}$ is a vector of unknown parameters. Assume $\boldsymbol{\theta}$ can be consistently estimated with $\widehat{\boldsymbol{\theta}}$. Before using $\widehat{\mathrm{M}} = \mathrm{M}(\widehat{\boldsymbol{\theta}})$ to address a practical issue, we need to be confident that M provides a *sufficiently accurate* approximation to F, where the accuracy is gauged relative to the practical issue. In other words, acknowledging Box's insight that "all models are wrong, but some are useful", how can we assess if M is useful?

To answer this question, graphical diagnostic methods are often employed to allow visualization of features that the data should exhibit if the model holds. Judging whether such features are present or absent in any particular diagnostic plot can be problematic. In this article we take a Bayesian sampling approach to aid in this task.

Bayesian Model Checking Diagnostics

Box (1980) proposed a Bayesian diagnostic for checking M based on the following. Conditional on M, the marginal, or predictive, distribution of $\boldsymbol{y}$ can be described by its density

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

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, 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

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

where the probability is calculated under M. A "small" value of α indicates that $\boldsymbol{y}_d$ would be unlikely to be 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, $g_i(\boldsymbol{y})$, 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.

(1) exists only for proper priors; Rubin (1984) proposed an alternative approach that can work with improper 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.

A Sampling Interpretation

Another way to think about the above approach is in terms of a sampling simulation. Draw a value of $\boldsymbol{\theta}$ from its prior distribution (for Box) or posterior distribution (for Rubin), 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 usefulness 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 prior (or posterior) distribution for $\boldsymbol{\theta}$. In particular, the methodology can be applied in situations where quantities such as (2) cannot be derived analytically. The choice of prior or posterior for generating realizations of $\boldsymbol{\theta}$ from M is discussed in Pardoe (2001).

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

Marginal Model Plots

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 some fixed h. Add a non-parametric mean estimate, say a cubic smoothing spline with fixed smoothing parameter, to the plot. The corresponding mean estimate under $\widehat{\mathrm{M}}$ can be obtained from a non-parametric mean estimate for the regression of the fitted values under $\widehat{\mathrm{M}}$ on h. Add this mean estimate to the plot 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 F allows point-wise comparison of the two estimates, since any estimation bias should cancel.

Ideas for selecting useful functions h to consider in practice include fitted values, individual predictors, and linear combinations of predictors. If M is a useful approximation to F, then for any quantity h the marginal mean estimates should agree. Any indication that the estimated marginal means do not agree for one particular h calls M into question; if they agree for a variety of plots, there is support for M.

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 M into question? Even if $\mathrm{M} = \mathrm{F}$, the estimated marginal means in a MMP would not match exactly. So, a technique is needed to visualize the variability in M to assess whether it would be reasonable for the data to be generated by such an M. The sampling interpretation for the Bayesian model checking diagnostics of Box and Rubin provides such a technique: for any particular MMP, calculate mean estimates for fitted values corresponding to individual samples from either the prior distribution (for Box) or posterior distribution (for Rubin) 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, $\mathrm{B}_t, t = 1, \ldots, m$, and obtain what we call a *Bayes marginal model plot* (BMMP) for the mean. Note that the $\boldsymbol{\theta}$ samples are being used directly here, as suggested above.

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 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 M provides a useful description of the conditional distribution of $y|\boldsymbol{x}$.

An Example

The Normal linear regression model can be written

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

where $\boldsymbol{X}$ is the design matrix, $\boldsymbol{\theta} = (\beta^T, \sigma^2)^T$, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters, σ^2 is the unknown error variance constant, and $\boldsymbol{W}$ is a known diagonal weight matrix.

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 σ^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 χ^2 random variable with $n-p$ degrees of freedom. Then, holding σ^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 $\boldsymbol{X}\boldsymbol{\beta}_t$. This procedure is straightforward to program in S-Plus using the `smooth.spline` function to obtain the non-parametric mean estimates with user-specified smoothing parameter—further details are available in Pardoe (2001).

Consider an example on maximizing yield, y, in a two-stage chemical process with temperatures (T_1, T_2) and times (t_1, t_2) of reaction at the two stages, and concentration (C) of one of the reactants at the first stage. Box (1954) based an analysis of these data on a full second-order response surface model. However, fitting 21 parameters with 32 observations may be somewhat ambitious, and Box notes that the data appear to support a response surface that is a two-dimensional ridge system embedded in the five-dimensional space of predictor variables. Cook (1998) suggests that a three-dimensional ridge system based on two linear combinations of the transformed predictors, lc_1 and lc_2, may in fact be sufficient. Accordingly, we fit a second-order model to lc_1 and lc_2; BMMP's for the mean in the directions of the fitted values for this model and lc_2, are shown in Figure 1.

Both BMMP's show the mean estimates under F lying inside the mean estimate bands under $\widehat{\mathrm{M}}$. There is little evidence in *these* plots to call this model into question. In fact, BMMP's for this 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 model.

The fitted model can then be used with confidence to address the practical question of maximizing yield. Because the model represents a three-dimensional ridge system, a range of predictor settings can maximize yield. Box (1954) discusses ways to identify such settings from the fitted model. For example, the experimenter may be interested mainly in settings for (T_1, T_2) that maximize yield in minimum time $t_1 + t_2$ and minimum concentration C.

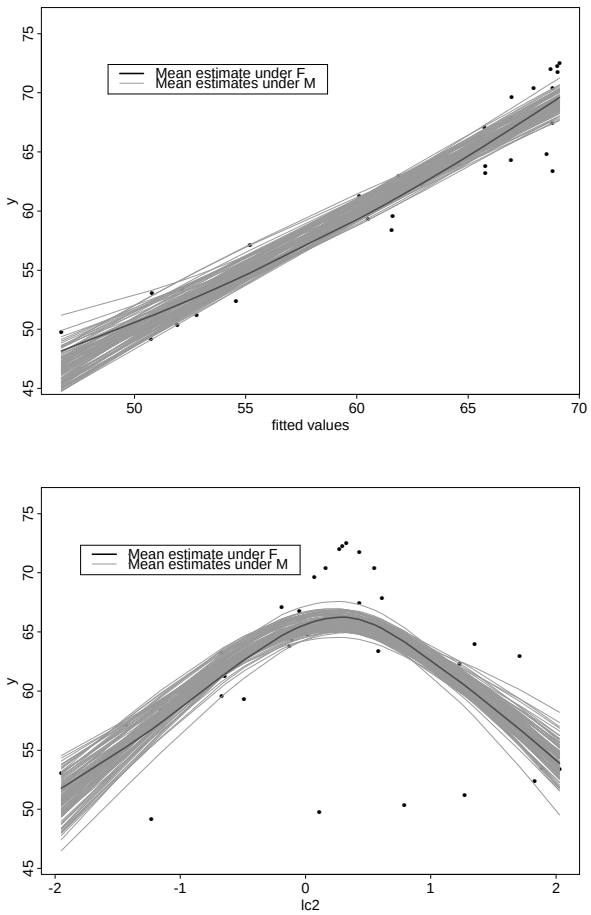

Figure 1: BMMP's for the mean.

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 usefulness of the model. Work is in progress to extend the methodology to variance function estimates, to generalized linear models, to other plots used in regression diagnostics, and to complementary quantitative methods.

References

Box, G. E. P. (1954). The exploration and exploitation of response surfaces: Some general considerations and examples. *Biometrics 10*, 16–60.

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.

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

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

Pardoe, I. (2001). A Bayesian sampling approach to regression model checking. *Journal of Computational and Graphical Statistics*. In press.

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