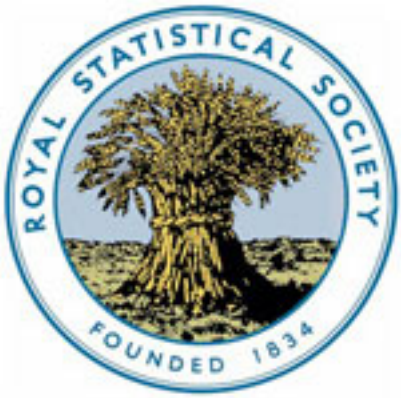

Appl. Statist. (2000)
49, *Part* 2, *pp.* 247–268

Diagnostic checks for discrete data regression models using posterior predictive simulations

Andrew Gelman

Columbia University, New York, USA

and Yuri Goegebeur, Francis Tuerlinckx and Iven Van Mechelen

Katholieke Universiteit Leuven, Belgium

[Received September 1997. Final revision July 1999]

Summary. Model checking with discrete data regressions can be difficult because the usual methods such as residual plots have complicated reference distributions that depend on the parameters in the model. Posterior predictive checks have been proposed as a Bayesian way to average the results of goodness-of-fit tests in the presence of uncertainty in estimation of the parameters. We try this approach using a variety of discrepancy variables for generalized linear models fitted to a historical data set on behavioural learning. We then discuss the general applicability of our findings in the context of a recent applied example on which we have worked. We find that the following discrepancy variables work well, in the sense of being easy to interpret and sensitive to important model failures: structured displays of the entire data set, general discrepancy variables based on plots of binned or smoothed residuals *versus* predictors and specific discrepancy variables created on the basis of the particular concerns arising in an application. Plots of binned residuals are especially easy to use because their predictive distributions under the model are sufficiently simple that model checks can often be made implicitly. The following discrepancy variables did not work well: scatterplots of latent residuals defined from an underlying continuous model and quantile–quantile plots of these residuals.

Keywords: Bayesian statistics; Binary regression; Generalized linear models; Quantile–quantile plots; Realized discrepancies; Residual plots; Sequential design; Stochastic learning models

1. Introduction

Model checking with discrete data regressions can be difficult because the usual methods such as residual plots have complicated reference distributions that depend on the parameters in the model (see Landwehr *et al.* (1984)). Posterior predictive checks have been proposed to deal with this problem as a Bayesian approach to classical goodness-of-fit testing (see Gelman *et al.* (1995, 1996) and Rubin (1984)). A key issue in setting up diagnostics is the choice of which aspects of the data to check. This is of particular concern when using posterior predictive checks because of the unlimited flexibility that they allow in choosing discrepancy variables. However, the main points of this paper—on choosing discrepancy variables and displaying their comparisons with the reference distribution—should also be relevant to related choices of reference distributions such as cross-validation and, in some contexts, prior predictive distributions (see Gelfand *et al.* (1992) and Weiss (1996)). As discussed by Gelfand

Address for correspondence: Andrew Gelman, Department of Statistics, Columbia University, New York, NY 10027, USA.
E-mail: gelman@stat.columbia.edu

 0035–9254/00/49247

(1996), a key issue in applied model checking, in any statistical framework, is being able to assess the model fit directly in a variety of ways.

In this paper, we consider several natural discrepancy variables in the context of a specific example that we introduce in Section 2. Although the methods described here are quite simple, we believe that they are not used as often as they could be, possibly because

(a) their theoretical justification has not always been clear and
(b) in the traditional non-Bayesian context it is not usual to work with discrepancy variables that depend on both data and parameters.

One purpose of this paper is, in fact, to demonstrate the simplicity and wide applicability of residual-type diagnostics for discrete data regression models, and also to explore useful methods of graphically displaying these checks. In Section 3 we review posterior predictive checking and in Section 4 we propose several classes of discrepancy variables for general discrete data regression problems as illustrated in the context of a specific logistic regression example. Section 5 introduces some model checks that are more specifically tailored to the problem under study. Section 6 briefly illustrates how plots of average residuals have been useful to us in another example of discrete data modelling. We conclude in Section 7 with a discussion and general recommendations.

2. A logistic regression model applied to a behavioural learning experiment

2.1. The experimental data and the historical context

We investigate the effectiveness of various model checks for an analysis of a logistic regression model applied to data from an experiment on behavioural learning. In this experiment, described in Bush and Mosteller (1955), each of 30 dogs was given a sequence of 25 trials; in each trial, a light was switched on for 10 s and then an electric shock was applied to the metal cage in which the dog was sitting. In each trial, the dog had an opportunity, once the light went on, to jump into an adjoining cage and thus to avoid the shock. In the initial trial, all the dogs received the shock (since they did not know the meaning of the signal), and in the succeeding 24 trials they learned to avoid it. Fig. 1(a) displays the experimental data for the 30 dogs, ordered by the time of the last trial in which they were shocked. (This ordering has nothing to do with the order in which the experiment was performed on the dogs; we choose it simply to make the between-dog variation in the data more visible.) Interest lies in the factors that affected the dogs' learning; in particular, did they learn more from successful avoidances than from shocks? Another question is, can the variation in responses among the 30 dogs be explained by a single stochastic learning model, or is there evidence in the data for underlying between-dog variation?

We choose this example to study model checking methods because the data and the associated stochastic learning model have an interesting structure, with replications over dogs and the probability of avoidance of an electric shock dependent on previous outcomes. As we shall see, the sequential nature of the model has important implications for some of the posterior predictive distributions that are used to calibrate the model checks. Specifically, the logistic regression model fits these data reasonably well but has interesting systematic patterns of misfit. These data are also of historical interest because the analysis by Bush and Mosteller (1955) included an early example of simulation-based model checking: they compared the observed data with simulations of 30 dogs from their model (with parameters fixed at their maximum likelihood estimates, which was reasonable in this case since they are accurately

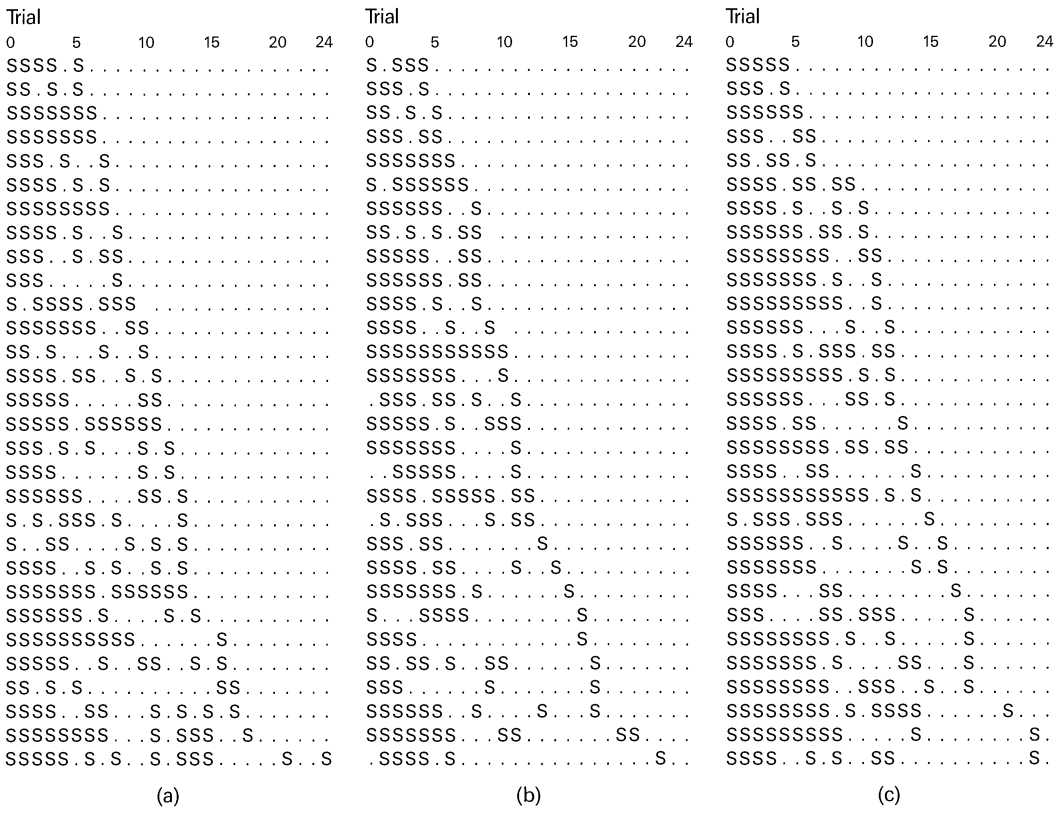

Fig. 1. (a) Sequence of shocks (S) and avoidances (.) for 25 trials on each of 30 dogs, from Bush and Mosteller (1955) (the dogs here are ordered by the time of the last shock, with ties broken randomly) and similar displays for 30 dogs simulated from (b) the logistic and (c) the logarithmic regression models conditionally on the estimated parameters for each model

estimated from this data set). Further simulation-based model checks for these data were considered by Sternberg (1963).

2.2. Setting up the logistic regression model and estimating its parameters

We use the notation $y_{jt} = 1$ or $y_{jt} = 0$ to indicate a shock or avoidance for trial t on dog j, for $t = 0, \ldots, 24$ and $j = 1, \ldots, 30$. We fit a logistic regression model,

$$\Pr(y_{jt} = 1|\beta) = \pi_{jt} = \text{logit}^{-1}(\beta_0 + \beta_1 X_{1jt} + \beta_2 X_{2jt}), \tag{1}$$

where

$$X_{1jt} = \sum_{k=0}^{t-1} (1 - y_{jk}) \tag{2a}$$

is the number of previous avoidances and

$$X_{2jt} = \sum_{k=0}^{t-1} y_{jk} \tag{2b}$$

is the number of previous shocks. On reflection, we can realize that this model is not ideal for these data. In particular, the experiment is designed so that all dogs receive shocks at trial 0 (which in fact happens, as is shown in Fig. 1(a)), whereas the logistic regression model is structured always to have a non-zero probability of both outcomes. This problem could be addressed in many ways, e.g. by fitting a logarithmic instead of a logistic link (the approach used by Bush and Mosteller (1955) and discussed further in Section 5) or by simply fitting the model to the data excluding trial 0.

To start, however, we fit the logistic regression model to the entire data set and examine what aspects of model misfit are uncovered by various posterior predictive checks. This is an interesting question because it is standard practice to fit a logistic regression model to binary data without seriously considering its appropriateness. (One reason for this is that for many problems there is no clearly preferable alternative to logistic regression, and simple tricks like changing the link function or discarding non-informative data are not generally sufficient to allow a simple model to fit.) In these cases of routine use, we would like to have routine model checks (by analogy with residual plots in normal regressions) that would give the user some idea of the model's problems. The goal of such methods is not to 'accept' or 'reject' a model but rather to highlight important areas where it does not fit the data.

For the Bayesian analysis, we assume an (improper) uniform prior distribution on the parameters $(\beta_0, \beta_1, \beta_2)$. We are comfortable with a non-informative prior distribution in this example because there are enough data to estimate the three parameters fairly accurately. Simulations from the posterior distribution were obtained from the Metropolis algorithm, using importance resampling draws from a mode-based normal approximation as starting-points for five independent sequences. Approximate convergence was achieved after 2000 iterations, in the sense that the potential scale reduction was less than 1.1 for all the parameters (see Gelman and Rubin (1992)). Taking the second half of all our simulated series yields 5000 draws of the parameter vector—i.e. a 5000×3 matrix. The posterior medians of β_0, β_1 and β_2 are 1.80, -0.35 and -0.21 respectively. The negative coefficients β_1 and β_2 imply that the probability of a shock declines after either a shock or an avoidance, with $|\beta_1| > |\beta_2|$ implying that avoidances have a larger effect.

3. Principles of posterior predictive checks

Monitoring the quality of a statistical model implies the detection of systematic differences between the model and observed data. Posterior predictive checks set this up by generating replicated data sets from the posterior predictive distribution of the statistical model under consideration; these replicated data sets are then compared with the observed data set with respect to any features of interest. The functions of data and model parameters that we use to compare with the model are called *test variables* or *discrepancy variables*; we also consider the special case of *test statistics*, which depend on the (replicated) data only.

To formalize these principles, we use the notation $y = (y_1, \ldots, y_n)$ for discrete observed data, X for the matrix of predictor variables and β for the vector of all parameters. We assume that a Bayesian model has been fitted and that we have a set of vectors β^l, $l = 1, \ldots, L$, drawn from the posterior distribution $p(\beta|X, y)$; see, for example, Dellaportas and Smith (1993), Albert and Chib (1993), Gelman *et al.* (1995) and Clayton (1996) on guidelines on how to perform the computations.

We further assume that, for each of these draws, a replicated data set $y^{\text{rep}l}$ has been simulated from the predictive distribution of the data $p(y^{\text{rep}}|X, \beta = \beta^l)$; the ensemble of simulated data sets $(y^{\text{rep}1}, \ldots, y^{\text{rep}L})$ thus represents the posterior predictive distribution

$p(y^{\text{rep}}|X, y)$. For simplicity, we suppress the conditioning on X in the notation that follows, but in some examples we shall allow X to vary and simulate X^{rep} as well (see Section 4.1). Posterior predictive simulation of the replicated data sets y^{rep}, conditionally on β, is usually extremely easy—typically requiring nothing more than simulation from known independent distributions—even though obtaining posterior simulations of β usually requires complicated Markov chain simulation methods.

We check the model by means of discrepancy variables $T(y, \beta)$. If β were known, we could perform a goodness-of-fit test by comparing the observed $T(y, \beta)$ with the distribution of the discrepancy variables in the replications, $T(y^{\text{rep}}, \beta)$, with the statistical significance of the test summarized by a p-value, $P = \Pr\{T(y^{\text{rep}}, \beta) > T(y, \beta)|y, \beta\}$. (Here, we consider only one-sided tests, with the understanding that the corresponding two-sided p-value is $2\min(p, 1-p)$.) In the more usual case that β is unknown, the test comparison is averaged over the uncertainty in β (i.e. the posterior distribution), with a posterior predictive p-value,

$$\Pr\{T(y^{\text{rep}}, \beta) > T(y, \beta)|y\} = \int \Pr\{T(y^{\text{rep}}, \beta) > T(y, \beta)|y, \beta\}\, p(\beta|y)\, \mathrm{d}\beta,$$

which can be estimated from the simulations by $\Sigma_{l=1}^{L} \mathbf{1}_{T(y^{\text{rep}l}, \beta^l) > T(y, \beta^l)}/L$, where $\mathbf{1}_A$ is the indicator function which is 1 if the condition A is true and 0 otherwise.

As to the choice of discrepancy variables, it is important to note that we focus here on methods for detecting systematic discrepancies between the model and data, not on the related problem of discovering outliers in otherwise reasonable models. Some of the discrepancy variables that we develop have been used in Bayesian methods for outlier detection (see Albert and Chib (1995), Chaloner and Brant (1988) and Chaloner (1991)) but there with a focus on individual observations rather than on larger patterns. By comparison, the discrepancy variables that we consider often average over sections of the data. In addition, we seek discrepancy variables that are easy to interpret and are also generally applicable to a wide range of problems. In many cases, this means that we would like to check qualitative features of the model (e.g. independence, monotonicity and unimodality) to give a better understanding of directions of model improvement.

The application of the posterior predictive check method goes as follows. Several discrepancy variables are chosen to reveal interesting features of the data or discrepancies between the model and the data. For each discrepancy variable, each simulated *realized value* $T(y, \beta^l)$ is compared with the corresponding simulated *replicated value* $T(y^{\text{rep}l}, \beta^l)$. Large and systematic differences between realized and replicated values indicate a misfit of the model to the data, in the sense that the observed data do not look typical, in this respect, of the data predicted under the model. In some cases, differences between the realized data and replications are apparent visually; other times, it can be useful to compute the p-value of a realized discrepancy to see whether it could plausibly have arisen by chance under the model.

4. General checks for discrete data regressions

In this section, we systematically characterize goodness-of-fit checks for discrete data regressions, illustrating throughout with the logistic regression model fit to the data from the dog experiment. We shall first discuss the issue of defining replications for the dog example. Next we shall use a few simple checks. In the final three subsections we shall discuss checks based on residuals.

4.1. Defining replications for the dog example

To perform model checks, the data must be compared with a reference distribution of possible replicated data sets. In a usual logistic regression model, this would be performed by fixing the matrix X of predictor variables and then, for each simulated parameter vector β^l, drawing the 25×30 responses $y_{jt}^{\text{rep}\,l}$ independently,

$$y_{jt}^{\text{rep}\,l} \sim \text{Bernoulli}\{\text{logit}^{-1}(X_{jt}\beta^l)\}, \tag{3}$$

to yield a simulated data set $y^{\text{rep}\,l}$. (The notation X_{jt} indicates the vector of predictors $(1, X_{1jt}, X_{2jt})$ defined in model (2).) Computing this for 5000 parameter vectors yields 5000 simulated data sets.

A stochastic learning model is more complicated, however, because the predictor variables X depend on previous outcomes y. Simulation of replicated data for a new dog must thus be performed sequentially. For each simulated parameter vector β^l: for each dog, $j = 1, \ldots, 30$, for trial $t = 0, \ldots, 24$,

(a) compute the vector of predictors, $X_{jt}^{\text{rep}\,l}$, based on the previous t trials for dog j, and
(b) simulate $y_{jt}^{\text{rep}\,l}$ as in model (3).

Interestingly, the two forms of reference distribution have the same likelihood for the observed data but different implications for the replications (since the sequential imputations alter X as well as y). Thus this aspect of the model is relevant to Bayesian model checking but not for Bayesian inference conditional on the model.

4.2. Simple checks for the dog example

Fig. 2 displays posterior predictive checks for two simple test statistics: the mean and standard deviation of the number of shocks per dog. In each plot, the observed value of $T(y)$ is shown as a vertical bar in a histogram representing 1000 draws of $T(y^{\text{rep}})$ from the posterior distribution. From Fig. 2(a), we see that the mean number of shocks is fitted well by the model. Fig. 2(b) shows that the observed standard deviation is slightly higher than expected under

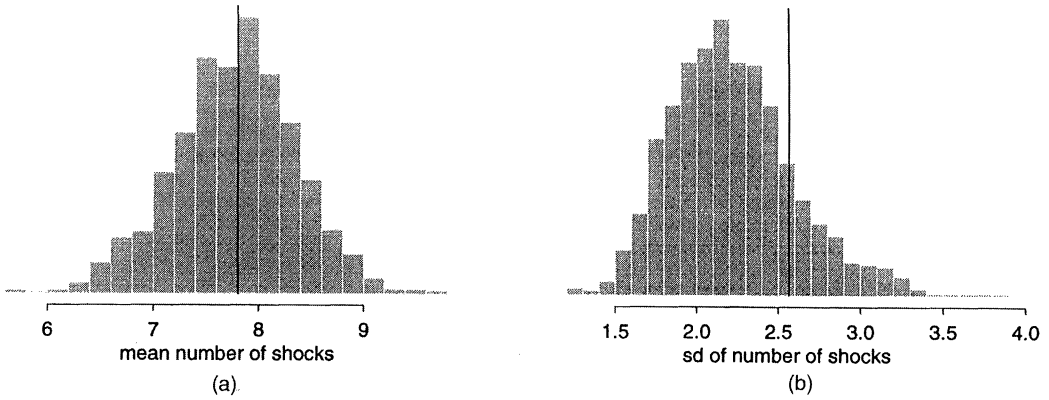

Fig. 2. Posterior predictive checks for (a) the mean and (b) the standard deviation of the number of shocks among the 30 dogs (the bars indicate the observed values of the test statistics $T(y)$ and the histograms display $T(y^{\text{rep}})$ from 1000 draws of y^{rep} under the logistic model)

the model, but the discrepancy is not statistically significant, i.e. we could expect to see such discrepancies occasionally just by chance, even if the model were correct.

We further illustrate the idea of posterior predictive checking with a visual comparison of the observed data with a replication under the assumed model. Fig. 1(a) shows the observed data with a single replicated data set y^{rep} (Fig. 1(b); ignore Fig. 1(c) for now). The visual display (aided by ordering the 30 dogs in each data set in order of the time of their last shock) shows some interesting differences between the real and simulated dogs. The 'discrepancy variable' here is simply the graphical display of the data, y. Strictly, a posterior predictive check should compare y with several draws of y^{rep}, but in this case a single draw is informative because of the internal replication of 30 independent dogs in a single data set.

The graphical comparison in Fig. 1 can be used to suggest more focused diagnostics. For example, the replicated dogs appear to have too few shocks in the early trials, compared with the real dogs. This is substantively relevant because the purpose of the model is to understand the learning behaviour of the dogs. To check this pattern more formally, we display in Fig. 3 the proportion of avoidances among the 30 dogs (i.e. $1 - \bar{y}_t$) *versus* time t. Overlain on the graph are the corresponding time series for 20 random draws $y^{\text{rep}l}$ from the posterior predictive distribution. Compared with the data, the model predicts too many avoidances in the first two trials and too slow an improvement in the first five trials. The model thus does not capture the rate of learning at the beginning of the experiment.

4.3. *Checks based on realized discrete residuals*

We now discuss more general checks for comparing discrete data with their predicted values. Perhaps the simplest approach, by analogy with continuous models, is to work with the vector of *realized residuals*, $r = y - E(y|X, \beta)$. The residuals are themselves uncertain because they depend on β, and so any function of the residuals (such as a plot of residuals *versus* expected values) can be viewed as a realized discrepancy, which can be compared with replications from the posterior predictive distribution as described at the end of the previous section. Unfortunately, for discrete data, plots of residuals can be difficult to interpret

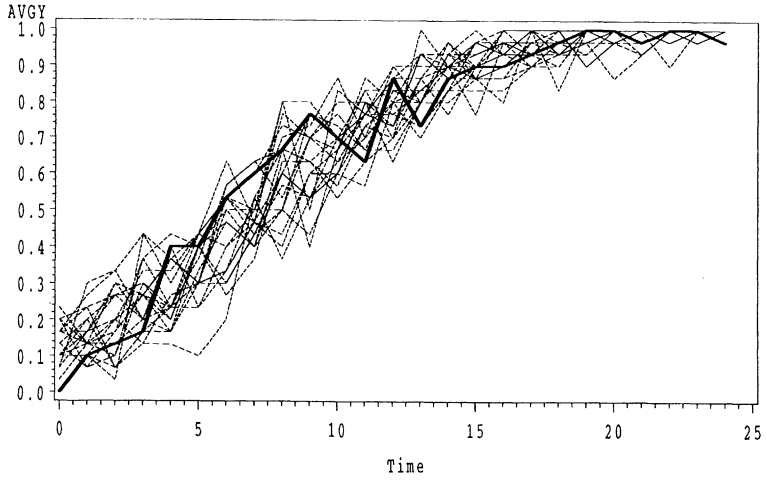

Fig. 3. Average number of avoidances among the 30 dogs, $1 - \bar{y}_t$, *versus* t: - - - - -, simulations from 20 replicated data sets under the logistic model

because, for any particular value of $E(y_i|X, \beta)$, the residual r_i can only take on certain discrete values; thus, even if the model is correct, the residuals will not generally be expected to be independent of predicted values or covariates in the model (see Albert and Chib (1995) for illustrations in a Bayesian context).

A standard way to make discrete residual plots more interpretable is to work with binned or smoothed residuals (see Landwehr *et al.* (1984)), which should be closer to symmetric about zero if enough residuals are included in each bin or smoothing category (since the expectation of each residual is by definition 0, the central limit theorem ensures that the distribution of averages of many residuals will be approximately symmetric). In particular, suppose that we would like to plot the vector of residuals r against some vector $w = (w_1, \ldots, w_n)$ that can in general be a function of X, β and perhaps y. We can bin the predictors and residuals by ordering the n values of w_i and sorting them into bins $k = 1, \ldots, K$, with approximately equal numbers of points n_k in each bin. For each bin, we then compute $\bar{w}_k$ and $\bar{r}_k$, the average values of w_i and r_i, respectively, for points i in bin k. The *binned residual plot* is the plot of the points $\bar{r}_k$ *versus* $\bar{w}_k$, which actually must be represented by several plots (which perhaps can be overlain) representing variability due to uncertainty of β in the posterior distribution.

Since we are viewing the plot as a test variable, it must be compared with the distribution of plots of $\bar{r}_k^{\text{rep}}$ *versus* $\bar{w}_k^{\text{rep}}$, where, for each simulation draw, the values of $\bar{r}_k^{\text{rep}}$ are computed by averaging the replicated residuals $r_i^{\text{rep}} = y_i^{\text{rep}} - E(y_i|X, \beta)$ for points i in bin k. In general, the values of w_i can depend on y (as in the lagged regression model that we are fitting to the dog data), and so the bins and the values of $\bar{w}_k^{\text{rep}}$ can vary between the replicated data sets.

Because we can compare with the distribution of simulated replications, the question arises, why do the binning at all? We do so because we want to understand the model misfits that we detect. Because of the discreteness of the data, the individual residuals r_i have asymmetric discrete distributions. The binned residuals are approximately symmetrically distributed. In general it is desirable for the posterior predictive reference distribution of a discrepancy variable to exhibit some simple features (in this case, independence and approximate normality of the $\bar{r}_k$) so that there is a clear interpretation of a misfit. This is, in fact, the same reason that we plot residuals, rather than data, against predicted values: it is easier to compare with an expected horizontal line than with an expected 45° line (see Tukey (1977) and Cleveland (1985)).

We apply the residual idea to the dog data by expressing the time series pattern of Fig. 3 in terms of residuals: we plot the average residuals at each time point, $\bar{r}_t = (1/30) \sum_{j=1}^{30} r_{jt}$, against t, where the residuals are computed from the logistic regression model: $r_{jt} = y_{jt} - \text{logit}^{-1}\{(X\beta)_{jt}\}$. A plot of $\bar{r}_t$ might be easier to interpret than the plot of $\bar{y}_t$ because, under the posterior predictive distribution, each $\bar{r}_t^{\text{rep}}$ has an expectation of 0, independent of the data for the previous t trials.

Because the average residuals $\bar{r}_t$ depend on β, the residual plot has posterior uncertainty. We display this, in Fig. 4, by overlaying plots of $\bar{r}_t^l$ *versus* t for 20 random draws β^l from the posterior distribution. (The 20 realizations are very similar because the parameters β are quite precisely estimated in the posterior distribution for this example.) The dotted curves on the plot show 95% bounds for $\bar{r}_t^{\text{rep}}$ under the model, calculated from the 5000 draws of (β, y^{rep}) from the posterior distribution for β and the predictive distribution for y^{rep}. In this posterior predictive distribution, the average residuals are independent with approximate normal distributions. Fig. 4 should be interpreted by comparing each of the 20 realizations separately with the posterior predictive distribution. The residuals show a clear pattern of too many shocks at the beginning and then too fast a reduction in the early trials. The model

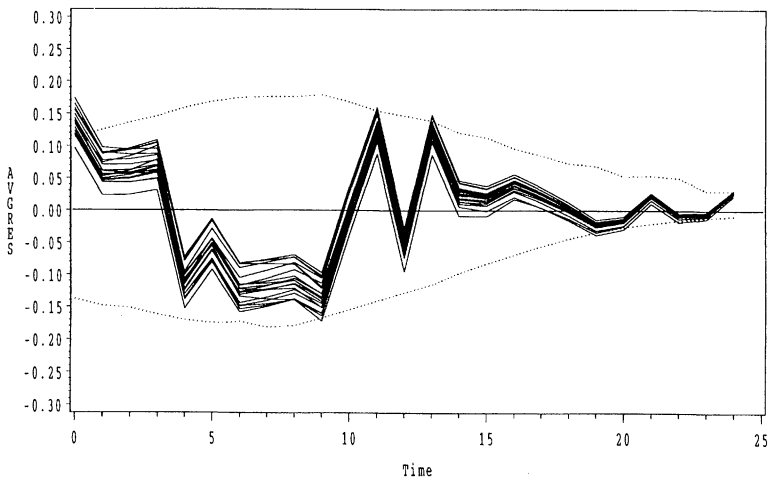

Fig. 4. Average discrete residuals $\bar{r}_t$ (averages over 30 dogs of $r_{jt} = y_{jt} - \text{logit}\{(X\beta)_{jt}\}$) *versus* time, from the observed data in Fig. 1(a): 20 plots of $\bar{r}_t$ are overlain, corresponding to 20 draws of β from the posterior distribution under the logistic model (·····, 95% predictive bounds under the model)

fits reasonably well (at least as far as the average residuals $\bar{r}_t$ are concerned) for the later trials.

Before going on, let us emphasize two points in the interpretation of Fig. 4. First, as usual with a residual plot, we are interested in finding patterns (such as dependence and linear or non-linear trends) that are not expected under the model. Thus it is possible for the plot to reveal problems with the model even if none of the individual points on the residual plot fall outside the 95% bounds. The dotted curves are displayed to give a visual impression of the scale of the variation among the posterior predictive replications. Second, the posterior predictive p-value of any particular residual can be computed by considering the p-value conditional on β, then averaging over the posterior distribution of β.

4.4. *Interlude: applying the procedure to simulated data*

As a comparison, we repeat the previous check, but applying it to simulated data (displayed in Fig. 1(b)) in place of the real data, i.e. we take the simulated data as if they were real, fit the logistic regression model, sample from the posterior distribution of β and simulate replications y^{rep}. We then plot $\bar{r}_t$ against t as in the previous section. The result is displayed in Fig. 5.

Originally, we performed this procedure for debugging—since the data in Fig. 1(b) are simulated from the model, they should probably show no problem in a model check. However, once the model was debugged, we noticed that Fig. 5 shows a striking pattern—a spike at trial $t = 7$. Suppose that we did not know that these data were simulated from the model—in that case, how should we interpret this spike? The natural first step would be to inspect the data and the experimental set-up to see whether there was some recording error or some change in the experiment at trial 7. If that is not the case, and if trial 7 is not of any particular scientific concern, it would probably be acceptable to show Fig. 5, with the implicit conclusion that a large residual at some point is explainable by random chance from the model.

Is our reasoning in Figs 4 and 5 consistent, or are we cheating by finding a problem with the fit to the real data but not to the simulated data? We do not believe that we are cheating because the two cases differ in an important aspect: in Fig. 4, the misfit is a consistent pattern

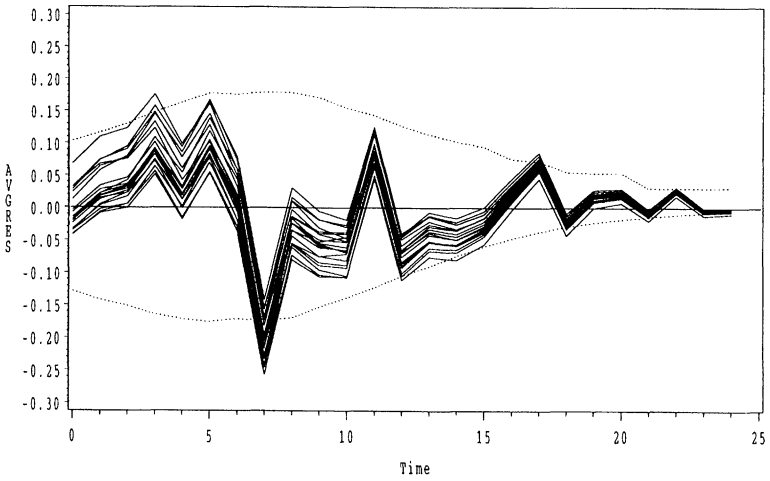

Fig. 5. Average discrete residuals $\bar{r}_t$ *versus* time, from the *simulated* data shown in Fig. 1(b): 20 plots of $\bar{r}_t$ correspond to 20 draws of β from the posterior distribution under the logistic model fit to the simulated data (because the replicated data have been simulated from the model, any patterns are due to chance alone; compare with Fig. 4)

that is relevant to our understanding of the rate of learning, a key objective of the study. In contrast, the potential misfit in Fig. 5, even if it happened to indicate a real problem, is isolated and does not affect that, on the whole, the model fits the data well.

4.5. General checks based on realized latent continuous residuals

A perhaps more appealing way to avoid the discreteness of the residuals is to work with residuals defined in an underlying continuous model, as suggested by Albert and Chib (1995). We discuss how this idea can be applied to posterior predictive checks, first in general terms and then in the context of the dog example.

Discrete data regression models can be formulated in terms of latent continuous variables. For example, the logistic regression model for binary data y, $\Pr(y_i = 1|\beta) = \text{logit}^{-1}\{(X\beta)_i\}$, can be interpreted in terms of a vector of latent continuous variables z_i with independent logistic distributions with scale 1 and locations $(X\beta)_i$, with the rule that $y_i = 1$ if $z_i > 0$ and $y_i = 0$ otherwise. The underlying continuous formulation is standard for probit models (in which case the variables z_i have normal distributions) and ordered multinomial distributions (in which case cut-points $c_0, c_1, \ldots$ must be estimated, so that, for example, $y_i = 0$ if $z_i < c_0$, $y_i = 1$ if $c_0 \leqslant z_i < c_1$, and so forth). In any of these models, the *realized continuous residuals* $\epsilon_i = z_i - E(z_i|X, \beta)$ are independent and identically distributed (with a known distribution $p(\epsilon)$ such as the standard normal or logistic distribution) under the model and for all β. As Albert and Chib (1995) discussed, the latent variables z_i are often simulated as a by-product of Gibbs sampler computations of the posterior distribution of β; when this is not the case, it is easy to draw simulations of z (and then to compute $\epsilon = z - X\beta$) conditionally on simulated β and the data y. The result is a set of posterior simulation draws ϵ^l, $l = 1, \ldots, L$, where each $\epsilon^l = (\epsilon_1^l, \ldots, \epsilon_n^l)$ is itself a vector.

We go beyond Albert and Chib (1995) by suggesting posterior predictive checks based on realized continuous residuals, thus making use of their simple and easy-to-understand reference distribution. When performing these checks, it is important to realize that, under the model, and *for any randomly chosen simulation draw l*, the realized continuous residuals

$(\epsilon_1^l, \ldots, \epsilon_n^l)$ should look like independent and identically distributed (IID) random draws from $p(\epsilon)$, since $\epsilon_1^{\text{repl}}, \ldots, \epsilon_n^{\text{repl}}$ are IID with distribution $p(\epsilon)$. Thus, for example,

(a) the residual plot of ϵ_i^l *versus* $(X\beta)_i$ (or, more generally, any function of X and β) should show no patterns, beyond what would be expected to occur by chance and
(b) the quantile–quantile (q–q)-plot of the ordered ϵ_i *versus* the quantiles of the theoretical error distribution (e.g. the logistic distribution) should look like a q–q-plot of random data (these, of course, are easy to simulate).

We emphasize that the IID nature of the realized continuous residuals refers to the set of n values of ϵ_i^l for a single l, *not* to the set of L random draws of ϵ_i^l for a single data point i. Plots of the posterior distribution of individual ϵ_i (i.e. collections of ϵ_i^l for $l = 1, \ldots, L$) can be useful for various purposes but are not to be expected to have the default distribution $p(\epsilon)$ (see Albert and Chib (1995) for more discussion of these sorts of plots). To see this, consider the very simple example in which $(X\beta)_i = 0$ for a particular data point i in a logistic regression model. In this case, if $y_i = 1$ or $y_i = 0$, ϵ_i has a logistic distribution truncated to be positive or negative respectively. Neither of these is the default distribution which in this case is an untruncated logistic distribution.

To see how this method might work in practice, we apply it to the dog example. Fig. 6 displays a plot of a single draw of the 750 continuous residuals ϵ_{jt} *versus* the linear predictor $(X\beta)_{jt}$, along with 95% bounds. The plot is disappointingly unrevealing because of the large posterior variation in each ϵ_{jt} (even conditional on β, the binary data just do not provide enough information about most of the underlying continuous parameters).

A quantile–quantile plot of the continuous residuals is also not informative. Fig. 7(a) displays 20 q–q-plots corresponding to 20 random draws of ϵ from the posterior distribution. Fig. 7(b) displays a similar plot from the predictive distribution of ϵ^{rep}. Figs 7(a) and 7(b) are indistinguishable, indicating that the distribution of the realized continuous residuals fits the model well (even though, as we have already learned, the model fits poorly in other ways).

Finally, we consider plots of averaged continuous residuals $\bar{\epsilon}_t = (1/30)\,\Sigma_{j=1}^{30}\,\epsilon_{jt}$ *versus* t. This is the analogy with Fig. 4 but using ϵ_{jt} instead of r_{jt}. Fig. 8 shows the result, which again

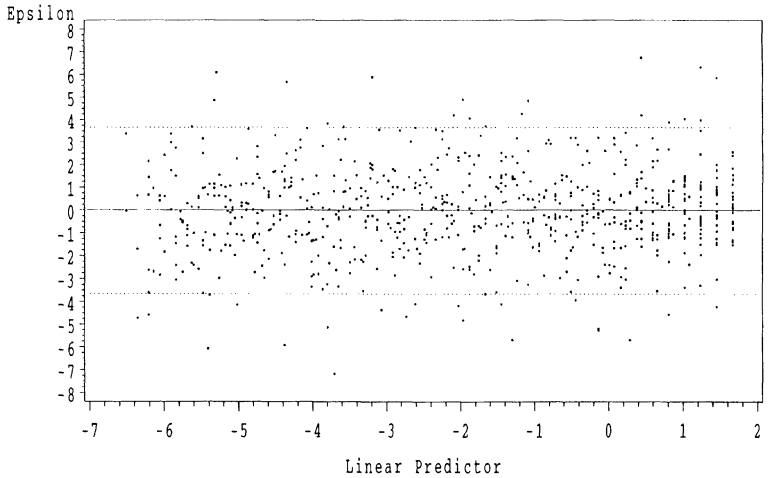

Fig. 6. Continuous residuals ϵ_{jt} *versus* linear predictor $(X\beta)_{jt}$ for a single draw from the posterior distribution of the logistic model: ·····, 95% predictive bounds under the model, in which ϵ_{jt} have IID logistic distributions

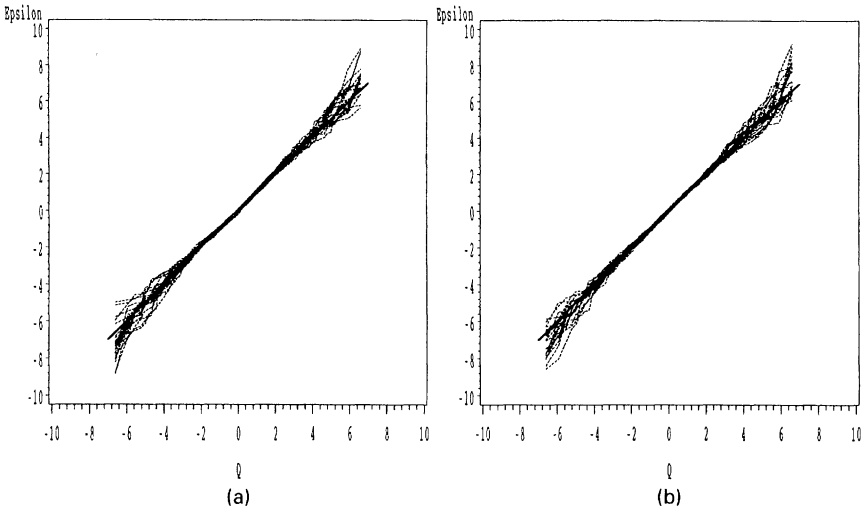

Fig. 7. (a) Quantile–quantile plots of 20 random draws of the vector of continuous residuals, ϵ_{jt}, from the logistic model and (b) posterior predictive distribution for (a), 20 quantile–quantile plots, each corresponding to a different set of 750 IID random draws from the logistic distribution

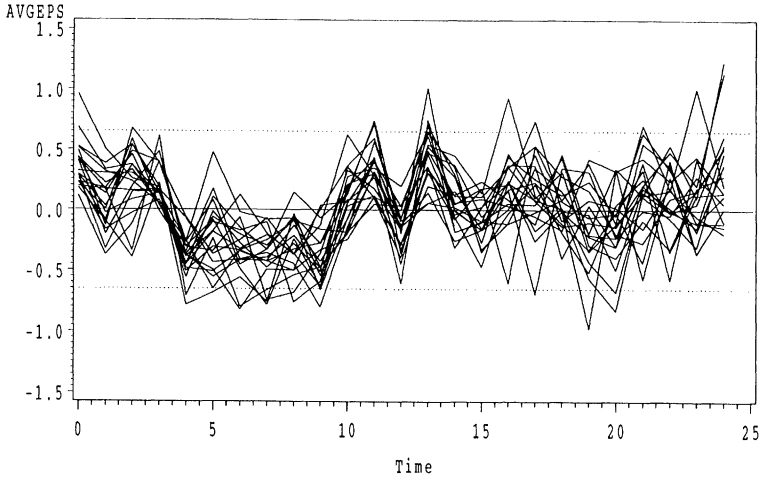

Fig. 8. Averaged continuous residuals $\bar{\epsilon}_t$ *versus* t for the logistic model: the 20 plots of $\bar{\epsilon}_t$ correspond to 20 draws of β from the posterior distribution: ·····, 95% predictive bounds from the model, computed from the distribution of the average of 30 IID draws from the logistic distribution

is disappointing. The model misfit in the early trials is visible, but much less clearly than in the plot of averaged discrete residuals in Fig. 4.

5. Developing checks for a particular discrete data regression problem

The previous section showed (in Figs 3 and 4) how we can diagnose a general lack of fit for the logistic model applied to the dog data. In this section, we move to a more reasonable logarithmic regression model for the same data (which was in fact fitted by Bush and Mosteller (1955)). Furthermore, we consider discrepancy variables tailored to the particular

problem under consideration. In particular, we develop targeted tests for examining evidence for between-dog variability in the data beyond that expected under this model. Our goal is both to explore the model fit in this example and to demonstrate how new test variables can be systematically developed in the context of a specific example.

5.1. Fitting the logarithmic regression model

We work with the following logarithmic regression model fitted by Bush and Mosteller (1955):

$$\Pr(y_{jt} = 1|\beta) = \pi_{jt} = \exp(\beta_1 X_{1jt} + \beta_2 X_{2jt}), \tag{4}$$

with X_{1jt} and X_{2jt} the number of previous avoidances and shocks respectively, as defined in expression (2). Unlike the logistic model (1), this model has no constant term because the probability of shock is fixed at 1 at the beginning of the experiment. In addition, β_1 and β_2 are restricted to be negative. We obtain posterior simulation draws of (β_1, β_2) from this model by using the Metropolis algorithm in an implementation that was very similar to that described at the end of Section 2.2. Fig. 9 displays posterior simulation draws for the parameters β_1 and β_2 in the model. The posterior median estimates for $(\exp(\beta_1), \exp(\beta_2))$ are (0.79, 0.92), indicating that an avoidance or a shock multiplies the predicted probability of shock by an estimated factor of 0.79 or 0.92 respectively.

Having fitted this improved model, we check its fit by using posterior predictive replications, which we simulate from the model as described in Section 4.1 (except by using the logarithmic rather than the logistic link). A single random draw from the posterior predictive distribution of 30 new dogs is displayed in Fig. 1(c). Formal checks, replicating Figs 3–8, show no apparent discrepancies with the model—the logarithmic link has fixed the problem with the early trials. We illustrate this in Figs 10 and 11, which replicate Figs 3 and 4 for the logarithmic model.

5.2. Test variables tailored to a specific question about the data

In any applied problem, it is appropriate to check aspects of the data and model that are of particular substantive interest. By their nature, such diagnostics can never be routine or automatic, but we can give some general suggestions. First, it is often useful to display the

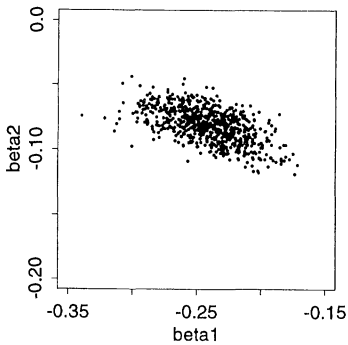

Fig. 9. Scatterplot of 2000 simulation draws of the parameters (β_1, β_2), which are the coefficients for the number of previous avoidances and the number of previous shocks respectively in the logarithmic regression model: β_1 is estimated to be more negative than β_2, indicating that avoidances have a larger effect than shocks in reducing the probability of future shocks

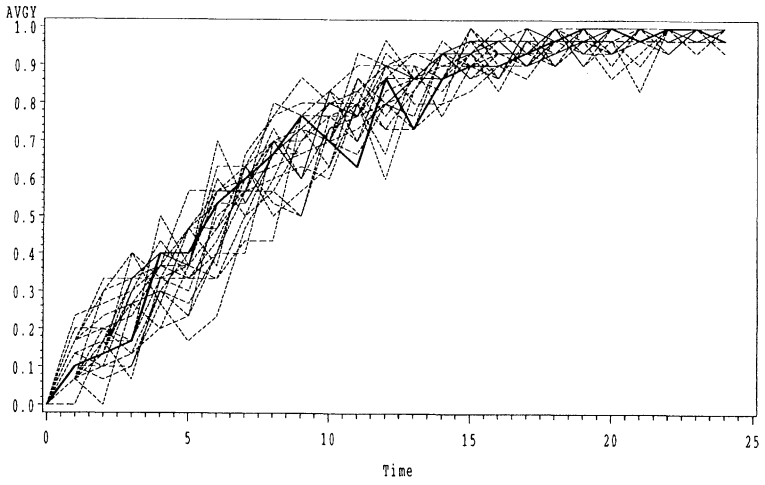

Fig. 10. Replication of Fig. 3 for the logarithmic model: ———, average number of avoidances among the 30 dogs, $1 - \bar{y}_t$, *versus* time; - - - - -, replications under the logarithmic model

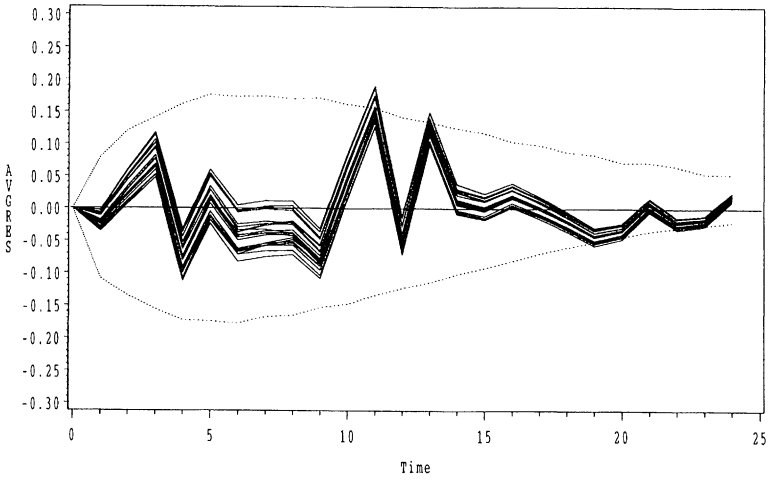

Fig. 11. Replication of Fig. 4 for the logarithmic model: the 20 curves plot the average discrete residuals $\bar{r}_t$ corresponding to 20 draws of β from the posterior distribution *versus* time (·····, 95% predictive bounds under the model)

entire data set (or, if that is not possible for a highly multivariate problem, various data rich summaries) and to compare with some posterior predictive replications of the data to obtain an idea of what would be expected under the model. Patterns seen in this sort of exploratory check can be used as the basis for more systematic model checks. As in exploratory data analysis in general, the reference distribution of replicated data sets provides a standard of comparison by which the observed discrepancies can be measured—the goal is not to find statistical significance but rather to reveal areas where the data look different from what would have been expected under the model. (But, as illustrated in Section 4.4, it can be perfectly reasonable to notice a possible problem in the exploratory analysis, to examine the situation more carefully and to conclude on substantive grounds that the observed discrepancy is not of scientific interest.) Second, we can directly compute the posterior

predictive distribution of any function of data and parameters by using the posterior simulations of (β, y^{rep})—and thus directly check the fit with respect to any easily computable discrepancy variable of interest. Third, it often makes sense to set up discrepancy variables with an eye on how the model might be improved—e.g. summarizing between-group variability if we are considering fitting a random-effects model. We now consider such tests in detail.

For the stochastic learning model, a key issue that has not been addressed so far is separating between-dog variability from stochastic learning. The data from the 30 dogs vary considerably. In the stochastic learning model this is explained by the fact that dogs learn more from avoidances than from shocks (i.e. $\beta_1 < \beta_2 < 0$) so a dog that is lucky at the beginning of the experiment is likely to perform well throughout, even in the absence of any real differences between dogs. However, we may consider an alternative explanation that the dogs indeed differ and, perhaps, there may be no additional learning associated with avoidances.

How can these data address this question? Most directly we could fit a random-effects model that includes both stochastic learning and between-dog variability. We consider this the best approach, but it is important to see what we can learn given the model that we have already fitted before undertaking a much more elaborate computation. Here we shall construct some discrepancy variables to check the fit of this stochastic learning model with respect to possible between-dog variability.

If the dogs truly vary in their underlying abilities, it should be possible to classify them by ability, on the basis of the data. Then, the higher ability dogs should perform better than predicted by the model and the lower ability dogs should perform worse. In this case, the classification of the dogs must depend on the responses themselves, since we have no other data to distinguish between the dogs. To be specific, we divide the dogs into two groups: the 'stupid' (or insensitive) dogs that are shocked in all the first five trials and the 'smart' (or sensitive) dogs that have at least one avoidance during this time. We then compare the two groups of dogs in various ways. (The number 5 was chosen to divide the dogs roughly into two equally sized groups.)

We began with the following simple test statistic: the average number of shocks after the first five trials for stupid dogs minus the average number of shocks after the first five trials for smart dogs:

$$T_1(y) = \frac{\sum_{\text{stupid dogs } j} \sum_{t=5}^{24} y_{jt}}{\sum_{\text{stupid dogs } j} 1} - \frac{\sum_{\text{smart dogs } j} \sum_{t=5}^{24} y_{jt}}{\sum_{\text{smart dogs } j} 1}. \tag{5}$$

If there is true between-dog variability, then we would expect the differences between stupid and smart dogs to be greater than expected by chance alone, and so the observed value of $T_1(y)$ would be larger than the values of $T_1(y^{\text{rep}})$ from posterior predictive replications simulated from the model.

In the observed data, stupid dogs are shocked an average of 4.8 times after the first five trials and smart dogs an average of 2.7 times, so $T_1(y) = 4.8 - 2.7 = 2.1$. Fig. 12 shows the histogram of 1000 simulated values of $T_1(y^{\text{rep}})$ under the posterior predictive distribution. Even under the assumptions of the model, T_1 is most likely to be positive; this is because, on the basis of the estimated parameters, a dog with more shocks in the past is more likely to be shocked in the future. The question is whether the dogs labelled stupid and smart differ in their future trials *even more* than expected on the basis of the model alone. (For each

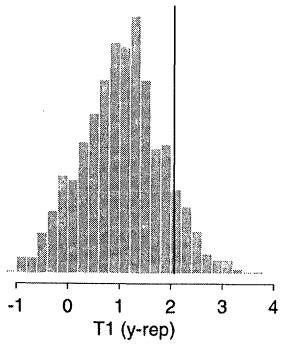

Fig. 12. Posterior predictive check for the test statistic $T_1(y)$, the difference between stupid and smart dogs in average number of shocks, defined in equation (5) in Section 5.2 (the bar indicates the observed value $T_1(y)$ and the histogram displays 1000 simulations of $T_1(y^{rep})$ under the logarithmic model)

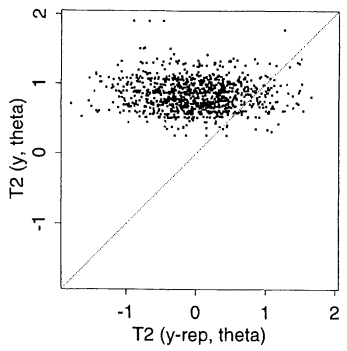

Fig. 13. Posterior predictive check for the discrepancy variable $T_2(y, \beta)$, the difference between stupid and smart dogs in average residuals, defined in equation (6) in Section 5.2 (the scatterplot displays 1000 simulations of $(T_2(y^{rep}, \beta), T_2(y, \beta))$ under the logarithmic model)

replication, $T_1(y^{rep})$ is computed with stupid and smart dogs defined on the basis of the first five trials of the simulated data set y^{rep}.) The observed value of 2.1 (indicated by a vertical line on the histogram) is about twice as high as would be expected under the model, but this deviation could be explained by chance (the p-value is 0.10).

For a slightly more sophisticated approach, we compared stupid and smart dogs on the basis of their average total *residuals*, $r_{jt} = y_{jt} - E(y_{jt}|\beta)$, after the first five trials:

$$T_2(y, \beta) = \frac{\sum_{\text{stupid dogs } j} \sum_{t=5}^{24} r_{jt}}{\sum_{\text{stupid dogs } j} 1} - \frac{\sum_{\text{smart dogs } j} \sum_{t=5}^{24} r_{jt}}{\sum_{\text{smart dogs } j} 1}. \tag{6}$$

This test variable is in some ways more directly interpretable than T_1 because it has an expected value of 0 in the posterior predictive replications since, if the model were true, the residuals have expectation 0 by definition. Fig. 13 shows a scatterplot of the realized and replicated values of the test variable. As expected, the replicated values are centred at zero, whereas the realized values are positive. The discrepancy is large (the stupid dogs receive nearly one shock more than the smart dogs, even after correcting for their expectations under the model) but the one-sided p-value of 0.08 indicates that this difference could be explained by chance.

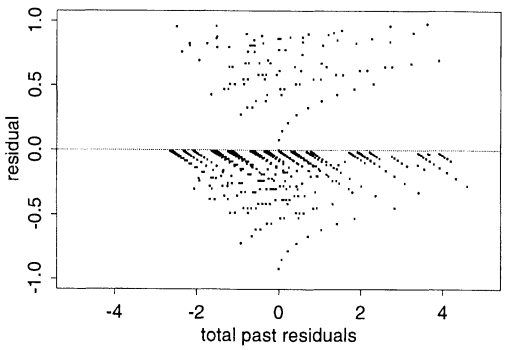

Fig. 14. Residuals *versus* sum of past residuals, based on the logarithmic model with the posterior median parameter estimates: each point represents one trial for one dog, with the residual for that trial plotted against the sum of the residuals for the previous trials for that dog; the discreteness of the data causes patterns in the residuals and makes the plot difficult to interpret

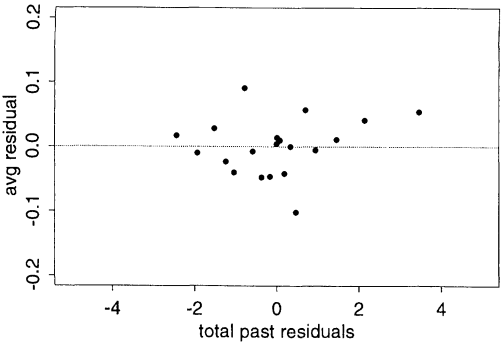

Fig. 15. Averaged residuals *versus* sum of past residuals, based on the logarithmic model with the posterior median parameter estimates: each point represents the average of one bin from the points in Fig. 14; by comparison, if the model were true, we would expect a patternless plot with zero mean

Although these results are not 'statistically significant', they are quite informative in that they give us a numerical understanding of the size of the between-dog variation that is apparent in the data. We have learned that it is certainly reasonable to suppose that the dogs truly vary in ability but such an assumption is not necessary to fit these data.

We also considered some more elaborate test statistics based on fitting separate regression models to the stupid and smart dogs and comparing the estimated regression coefficients. These tests did not seem particularly informative, which is consistent with experience in the psychology literature on individual differences about the lack of reliability of intraindividual correlations.

5.3. A residual plot tailored to a specific question about the data

The test variables described in the previous section are informative but are perhaps difficult to generalize because of the specific nature of the labelling of stupid and smart dogs. To address this problem, we constructed a new test variable, based on a residual plot, to assess between-dog variability. The idea is as follows: if the dogs truly vary more than would be expected under the model, then we would expect the residuals to be positively correlated within dogs, i.e., if the early residuals for a dog are positive (indicating more shocks than expected under

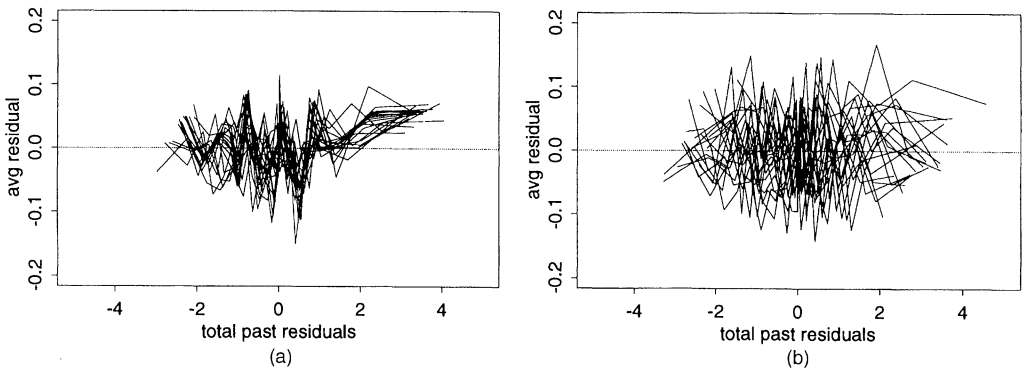

Fig. 16. (a) Averaged residuals *versus* sum of past residuals, based on the logarithmic model with 20 curves representing 20 random draws from the posterior distribution, and (b) 20 draws from the posterior predictive distribution if the model were true

the model), then we would expect the later residuals to be positive also. We check this by plotting the residual for each trial against the sum of the residuals from the previous trial for that dog. The plot, displayed in Fig. 14, shows distracting patterns caused by the discreteness of the data. The averaged binned residuals are plotted in Fig. 15 and appear to show some between-dog correlation: positive previous residuals are predictive of positive residuals on the current trial, in contradiction to the model's implicit prediction of mean zero residuals conditional on previous data.

For both Fig. 14 and Fig. 15, the residuals are computed with respect to the expected values based on the posterior median estimates of (β_1, β_2). We account for posterior uncertainty in the parameters in Fig. 16(a) by overlaying 20 lines, each corresponding to an averaged residual plot as displayed in Fig. 15, but with residuals computed with respect to a different random draw from the posterior distribution of (β_1, β_2). Fig. 16(b) displays corresponding simulations based on the posterior predictive replications. In the general language of model checking, Fig. 16(a) displays 20 draws of the realized discrepancy $T(y, \beta)$ and Fig. 16(b) displays 20 draws of the predictive discrepancy $T(y^{\text{rep}}, \beta)$. The comparison shows that the realized discrepancies are indeed different from average—if a dog had more previous shocks than expected in previous trials, it is more likely to be shocked than is predicted under the model—but this discrepancy is not larger than might be expected by chance alone given the relatively small size of the data set. Our substantive conclusion with this residual plot is thus similar to that obtained with the specially tailored test variable described in Section 5.2.

6. An ordered multinomial model in a study of pain relief

In our detailed analysis of the dog example, we found that the most useful discrepancy variables were

(a) general views of the data and checks tailored to particular issues specific to the model and the problem being studied and
(b) binned realized residuals plotted against predictor variables.

In fact, discrepancy variables of type (b) have been generally effective in many problems of discrete data analysis on which we have been involved recently; we illustrate with an example here.

Sheiner *et al.* (1997) described an analysis of a study of a pain relief drug in which the measured outcome, subjective pain relief of subjects in the trial, is measured on a 0–4 scale (from 'no pain relief' to 'complete pain relief'), with each subject j reporting pain relief y_{jt} at several time points t in the study, for a total of $n = 1500$ measurements. A pharmacokinetic model is fitted, modelling an underlying continuous pain relief level $\omega_j(t)$, with four continuous cut-off parameters $\eta_0, \ldots, \eta_3$ estimated, so that y_{jt} equals 0 if $\omega_j(t) < \eta_0$, 1 if $\eta_0 \leqslant \omega_j(t) < \eta_1, \ldots,$ 4 if $\omega_j(t) > \eta_3$, as in McCullagh (1980). This set-up is more complicated than a generalized linear model, but it still can be summarized as a vector $y = (y_1, \ldots, y_n)$ of observed data (stringing together the data y_{jt} from all subjects j and all time points t), a vector of parameters β and a model of independent outcomes: for each $i = 1, \ldots, n$,

$$\Pr(y_i = u|\beta) = \pi_{ui}(\beta), \qquad \text{for } u = 0, \ldots, 4. \tag{7}$$

Sheiner *et al.* (1997) fitted their model by using maximum likelihood, resulting in a point estimate of the parameters β and thus a point estimate of the matrix of data probabilities π_{ui}, $u = 1, \ldots, 4$, $i = 1, \ldots, n$. Because we have no other information, we shall treat the maximum likelihood estimate as if it represents the entire posterior distribution, which should give us conservative tests, as discussed in Section 7.2 later.

To check the model, we need the data vector y and the matrix of probabilities π, but *no other knowledge is required of the underlying model*, once we have been given the probabilities π. In this as in many examples, the model is difficult to fit, but it is easy to check once the fitting has been done. Fig. 17 displays a plot of binned discrete residuals *versus* predicted values, as described in Section 4.3. The lines show 95% bounds for the average residuals under the model, created on the basis of the multinomial distributions (7). In general, these would have to be computed by simulation, but in this case the large number of measurements per bin allows us to use the normal approximation based on the mean and variance of the averages of the multinomial distributions.

Fig. 17 is a good illustration because it is simple to create and reveals an interesting problem with the fit of the model: when the predicted values are low (i.e. predicting little to no pain relief), the actual pain relief is even less, and, when the predicted pain relief is high, the actual pain relief is even higher. However, the magnitude of the misfit is small, as can be seen by comparing the scales on the horizontal and vertical axes.

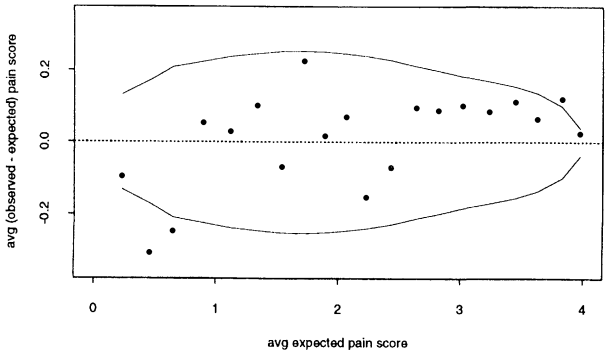

Fig. 17. Averaged residuals *versus* expected pain scores, with responses divided into 20 equally sized bins defined by ranges of expected pain scores, for the example in Section 6: the prediction errors are relatively small but with a consistent pattern that low predictions are too low and high predictions are too high (adapted from Gelman and Bois (1997) (·····, 95% bounds under the model)

In interpreting scatterplots such as Figs 14–17, it is important to remember that we usually are interested in patterns as well as individual points. For example, even if none of the points in Fig. 17 had fallen outside the 95% predictive error bounds, the plot would still indicate a model misfit because it shows a pattern that would not be expected if the model were appropriate.

7. Discussion

The flexibility of posterior predictive checks allows us to consider any discrepancy measure of interest. We considered several kinds of discrepancies on the basis of the following criteria: relevance to the particular problem being studied, simplicity of display and ability to check qualitative features of the model such as independence. The second and third criteria are satisfied by discrepancy measures with relatively simple posterior predictive distributions. In particular, residuals, by definition, have expectation 0; binned residuals have approximate normal distributions; in regression models with independent errors, the realized residuals are independent. We consider binned residual plots for residuals and also latent residuals.

7.1. What have we learned about posterior predictive checking from the dog example?

For the dog learning example, we tried a wide range of discrepancies. We were not surprised to find that comparing a structured display of the entire data set with a replication under the model (Fig. 1) was informative. More focused checks based on a particular question about between-dog variability (Figs 12 and 13) were also useful. Among the general residual plots, we found plots of binned realized residuals (Figs 4, 15, 16 and 17) to be sensitive to important model failures. Unfortunately, the plots based on latent continuous residuals (Figs 6–8) were too noisy to be useful in this example. The difficulty was that the discrete data provide very little information about the individual latent variables, and so any patterns in the latent residuals disappear in the posterior uncertainty.

How did we use the model checks in the dog learning example? We learned that the logistic regression model fits reasonably well in the later part of the experiment but not in the first few trials. This problem was fixed by the logarithmic regression model. We also learned that there is evidence for some between-dog variability but not necessarily more than what might be observed under the model. These conclusions illustrate an important general way of using these discrepancies: not as 'accept or reject' but as a way to understand the limitations of the model for future use, which is formalized by the model for replications y^{rep}.

We have found in general that binned realized residual plots are a useful adjunct to Bayesian regression-type models, and we illustrated with another example (Fig. 17) how such a plot can reveal interesting patterns of model misfit, even in an example in which we are not particularly familiar and applied the realized residual plot in an automatic fashion. We recommend that these plots be routinely used in regression modelling and imputation. In addition, it makes sense to examine data displays and also more focused discrepancies tailored to the particular problems under study. Just as model building is inherently flexible, especially in a modern computational context (see, for example, Smith and Roberts (1993) and Hastie and Tibshirani (1990) for discussions in the Bayesian and non-Bayesian contexts respectively), so can predictive checking be a versatile tool for understanding the fit of a model.

7.2. Relationship to other methods of predictive model checking

All model checking (and, one might argue, most exploratory data analysis; see Tukey (1977)) is based on a comparison of observed data with predictions; when considering a stochastic model, these predictions will be uncertain and we are thus comparing data with predictive distributions.

For us, posterior predictive checking is a natural approach since, from a Bayesian perspective, inference about the hypothetical replicated data y^{rep}, like all other unobserved quantities, can be summarized by a posterior distribution. However, other methods are available for defining reference distributions, and we suspect that the ideas in this paper about choosing test quantities will be relevant there also.

In the case that a full Bayesian model has not been fitted, and all that is available is an estimate, $\hat{\beta}$, we can consider classical tests that are equivalent to our approach but with the posterior distribution replaced by a point mass at this estimate. (This procedure is sometimes called a parametric bootstrap—see Efron and Tibshirani (1993).) Thus, the replicated data sets are simulated from $p(y^{\text{rep}}|\beta = \hat{\beta})$, which approximates the full posterior predictive distribution. In general, we prefer to work with the posterior distribution, but in many applications we can perform approximate model checks by using a reasonably accurate point estimate (for example, see Section 6). Posterior predictive tests can be viewed as the Bayesian extension of classical tests based on point estimates (see Gelman *et al.* (1996)) and in fact converge to the classical tests in the limit of large sample size. For finite samples, we expect tests based on the maximum likelihood estimate to tend towards conservatism, because the data should fit better to the best-fit point estimate than to a random draw from the posterior distribution.

Goodness-of-fit tests, both classical and Bayesian, can sometimes be difficult to interpret because the same data are used for both fitting and testing the model (see Gelman *et al.* (1996) and Meng (1994)). An alternative approach is a comparison with new data or cross-validation by splitting existing data. Cross-validation has its own problems (notably, that the inferences being tested are not the same as the final inferences constructed from all the data) but in many settings is an extremely useful and convincing procedure (see Gelfand *et al.* (1992) and Draper (1996) for discussions in the Bayesian context and Price *et al.* (1996) for an example from our own research). In any case, these methods are also forms of predictive checking and as such require test variables or summaries of the data. The concerns of relevance and effectiveness of the tests still apply, and so we believe that the lessons we have learned from our examples will be relevant to these other predictive model checking approaches.

In conclusion, the flexibility of modern statistical modelling and computation allows a corresponding flexibility in model checking, but this can only be fully exercised when we have an understanding of what model checks, and what display methods, are effective in practice. The analyses of the examples in this paper show how, in the tradition of exploratory data analysis, generally applicable model checks can reveal hidden features in discrete data.

Acknowledgements

We thank Xiao-Li Meng, the Editors and the referees for helpful comments. This work was supported in part by Fellowship F/96/9 and grant OT/96/10 of the Research Council of Katholieke Universiteit Leuven and Young Investigator Award DMS-9457824 of the US National Science Foundation.

References

Albert, J. H. and Chib, S. (1993) Bayesian analysis of binary and polychotomous response data. *J. Am. Statist. Ass.*, **88**, 669–679.

——— (1995) Bayesian residual analysis for binary response regression models. *Biometrika*, **82**, 747–759.

Bush, R. R. and Mosteller, F. (1955) *Stochastic Models for Learning*, ch. 11. New York: Wiley.

Chaloner, K. (1991) Bayesian residual analysis in the presence of censoring. *Biometrika*, **78**, 637–644.

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

Clayton, D. (1996) Generalized linear mixed models. In *Practical Markov Chain Monte Carlo* (eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter), pp. 275–301. London: Chapman and Hall.

Cleveland, W. S. (1985) *The Elements of Graphing Data*. Monterey: Wadsworth.

Dellaportas, P. and Smith, A. F. M. (1993) Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling. *Appl. Statist.*, **42**, 443–459.

Draper, D. (1996) Utility, sensitivity analysis, and cross-validation in Bayesian model checking. *Statist. Sin.*, **6**, 760–767.

Efron, B. and Tibshirani, R. J. (1993) *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Gelfand, A. E. (1996) Model determination using sampling-based methods. In *Practical Markov Chain Monte Carlo* (eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter), pp. 145–161. London: Chapman and Hall.

Gelfand, A. E., Dey, D. K. and Chang, H. (1992) Model determination using predictive distributions with implementation via sampling-based methods (with discussion). In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 147–167. New York: Oxford University Press.

Gelman, A. and Bois, F. Y. (1997) Discussion on 'Analysis of non-randomly censored ordered categorical longitudinal data from analgesic trials' (by L. B. Sheiner, S. L. Beal and A. Dunne). *J. Am. Statist. Ass.*, **92**.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (1995) *Bayesian Data Analysis*. London: Chapman and Hall.

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

Gelman, A. and Rubin, D. B. (1992) Inference from iterative simulation using multiple sequences (with discussion). *Statist. Sci.*, **7**, 457–511.

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

Landwehr, J. M., Pregibon, D. and Shoemaker, A. C. (1984) Graphical methods for assessing logistic regression models (with discussion). *J. Am. Statist. Ass.*, **79**, 61–71.

McCullagh, P. (1980) Regression models for ordinal data (with discussion). *J. R. Statist. Soc.* B, **42**, 109–142.

Meng, X. L. (1994) Posterior predictive p-values. *Ann. Statist.*, **22**, 1142–1160.

Price, P. N., Nero, A. V. and Gelman, A. (1996) Bayesian prediction of mean indoor radon concentrations for Minnesota counties. *Hlth Phys.*, **71**, 922–936.

Rubin, D. B. (1984) Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.*, **12**, 1151–1172.

Sheiner, L. B., Beal, S. L. and Dunne, A. (1997) Analysis of non-randomly censored ordered categorical longitudinal data from analgesic trials (with discussion). *J. Am. Statist. Ass.*, **92**.

Smith, A. F. M. and Roberts, G. O. (1993) Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. R. Statist. Soc.* B, **55**, 3–23.

Sternberg, S. (1963) Stochastic learning theory. In *Handbook of Mathematical Psychology* (eds R. D. Luce, R. R. Bush and E. Galanter), vol. 2, pp. 1–120.

Tukey, J. W. (1977) *Exploratory Data Analysis*. New York: Addison-Wesley.

Weiss, R. E. (1996) Bayesian model checking with applications to hierarchical models. *Technical Report*. Department of Biostatistics, University of California, Los Angeles.