Biometrika (1995), **82**, 4, *pp.* 747–59
Printed in Great Britain

Bayesian residual analysis for binary response regression models

BY JIM ALBERT

Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, Ohio 43403, U.S.A.

AND SIDDHARTHA CHIB

Olin School of Business, Washington University, St. Louis, Missouri 63130, U.S.A.

SUMMARY

In a binary response regression model, classical residuals are difficult to define and interpret due to the discrete nature of the response variable. In contrast, Bayesian residuals have continuous-valued posterior distributions which can be graphed to learn about outlying observations. Two definitions of Bayesian residuals are proposed for binary regression data. Plots of the posterior distributions of the basic 'observed – fitted' residuals can be helpful in outlier detection. Alternatively, the notion of a tolerance random variable can be used to define latent data residuals that are functions of the tolerance random variables and the parameters. In the probit setting, these residuals are attractive in that a priori they are a sample from a standard normal distribution, and therefore the corresponding posterior distributions are easy to interpret. These residual definitions are illustrated in examples and contrasted with classical outlier detection methods for binary data.

Some key words: Data augmentation; Gibbs sampling; Latent data; Outlier; Simulation.

1. INTRODUCTION

This paper focuses on the problem of outlier detection in discrete data regression models. Although the results presented are developed in the context of cross-sectional and longitudinal binary data, the ideas are more generally applicable to binomial outcomes. The canonical model that is analysed involves a set of independent variables $y = (y_1, \ldots, y_N)$, where y_i is a binary (0, 1) variable, such that

$$p_i := \text{pr}(y_i = 1) = F(x_i^{\mathrm{T}}\beta),$$

where $F(.)$ is a known cumulative distribution function, $x_i = (x_{i1}, \ldots, x_{ik})^{\mathrm{T}}$ are k measured covariates, and $\beta = (\beta_1, \ldots, \beta_k)^{\mathrm{T}}$ is an unknown vector of parameters. The problem of interest is the detection of outliers in the collection y. Later in the paper, this same problem is taken up for longitudinal discrete data models with random effects.

From a frequentist viewpoint, outlier detection in these models is based on the difference $y_i - \hat{p}_i$, where $\hat{p}_i = F(x_i^{\mathrm{T}}\hat{\beta})$ is the fitted probability for the ith observation, and $\hat{\beta}$ is the maximum likelihood estimate of β (Pregibon, 1981; Collett, 1991, Ch. 5). Specifically, one approach to detecting outliers is based on the set of Pearson residuals

$$\frac{y_i - \hat{p}_i}{\{\hat{p}_i(1-\hat{p}_i)\}^{\frac{1}{2}}},$$

and another on the deviance residuals

$$d_i = \text{sgn}\,(y_i - \hat{p}_i)\left\{2y_i \log\left(\frac{y_i}{\hat{p}_i}\right) + 2(1-y_i)\log\left(\frac{1-y_i}{1-\hat{p}_i}\right)\right\}^{\frac{1}{2}}.$$

These residuals can be useful in outlier detection for binomial data: see Duffy (1990) for the definitions relevant in that case. However, in the context of binary data the Pearson residual and the deviance residual have unknown sampling distributions and so the residual plots can be difficult to interpret.

Owing to the inadequacy of the use of these residuals for binary regression data, alternative methods have been suggested. In particular, authors have proposed the use of cross-validation probabilities (Davison, 1988), a simulated envelope about the half-normal plot of the deviance residuals (Atkinson, 1981), and a residual smoothing algorithm (Fowlkes, 1987). Although these techniques are useful, they do not address the basic problem of interpreting the size of the residuals.

In this paper, the outlier detection problem for binary regression is addressed from a Bayesian perspective. Zellner (1975), Zellner & Moulton (1985), Chaloner & Brant (1988) and Chaloner (1991) proposed and illustrated Bayesian definitions of residuals and this material is briefly reviewed in § 2·1 in the context of linear regression models. This literature motivates the consideration of two Bayesian residuals in § 2·2 that are useful in the binary data setting. The residuals and accompanying plots are demonstrated with a small simulated data set in § 3 and with two real data sets in § 4. It is shown that the posterior distributions of these residuals can communicate information that is not evident in classical residual plots. In § 5, the work is reviewed and other means of summarising the residual distributions are discussed.

2. BAYESIAN RESIDUALS

2·1. *Background*

In the normal linear model setting, Chaloner & Brant (1988) proposed a simple approach for the detection of outliers. Let the regression model be given as $Z_i = x_i^{\mathrm{T}}\beta + \varepsilon_i$, where the ε_i are a random sample from $N(0, \sigma^2)$, and $Z_i \in (-\infty, \infty)$ is observed. Given Z_i, consider the 'realised residual' $\varepsilon_i(\beta) = Z_i - x_i^{\mathrm{T}}\beta$. A priori, this residual is distributed $N(0, \sigma^2)$, and the posterior distribution of $\varepsilon(\beta) = (\varepsilon_1(\beta), \ldots, \varepsilon_M(\beta))^{\mathrm{T}}$ follows from the posterior distribution of β and σ^2. From this posterior distribution, the ith observation can be considered to be outlying if the posterior distribution of $\varepsilon_i(\beta)$ is located far from zero. Specifically, one can regard the ith observation as outlying if the posterior probability $\text{pr}\,(|\varepsilon_i| > K\sigma | Z)$ is large, where $Z = (Z_1, \ldots, Z_N)$.

Chaloner & Brant (1988) suggest choosing a value of K such that the prior probability of finding any outliers among the N observations is small, say 0·05. Alternatively, one can use a fixed value of K, say 2, and compare the prior and posterior probabilities that each individual residual exceeds $K\sigma$ in absolute value.

2·2. *Binary response residuals*

For the binary regression setting, consider comparing the observed binary observation y_i with the probability $p_i = F(x_i^{\mathrm{T}}\beta)$. As p_i is unobserved, the classical approach is to compare y_i with the ith fitted probability $\hat{p}_i$. However, the difference $y_i - \hat{p}_i$ is difficult to interpret since the reference distribution is over the sampling distribution of $y_i - \hat{p}_i$ which

is not known due to the binary response variable. On the other hand, from a Bayesian perspective, the parametric residual $r_i = y_i - p_i$ has a continuous-valued posterior distribution which can give information about outliers (Albert & Chib, 1993).

Note that r_i is a function of the regression vector β. Therefore, the precision of the knowledge about β, as reflected in its posterior distribution, will be reflected in the precision of the sizes of the residuals. Specifically, if the posterior distribution of p_i and the value of y_i are in conflict, then the posterior distribution of r_i will be concentrated towards extreme values. Since the support of the posterior distribution of r_i is on the interval $(y_i - 1, y_i)$, an observation $y_i = 0$ will be outlying if the posterior of r_i is concentrated towards the endpoint -1, and an observation $y_i = 1$ is unusual if the posterior of r_i is concentrated towards the value 1.

Albert & Chib (1993) and Dellaportas & Smith (1993) describe how to obtain a simulated sample of size G from the posterior distribution of β. Let the sample be denoted by $\{\beta^{(g)}\}_{g=1}^{G}$. Then the posterior distribution of p_i, and hence that of r_i, can be summarised from the simulated values $y_i - F(x_i^{\mathrm{T}}\beta^{(g)})$ $(g \leqslant G)$. In particular, summary values such as the median, quartiles, and the 5th and 95th percentiles can be obtained from the simulated values. One can see if a particular residual is unusually large by plotting the quantiles of the distribution of r_i against the posterior mean of the probability p_i.

An alternative approach to defining a Bayesian residual is based on the concept of a tolerance random variable from bioassay. Corresponding to each binary observation y_i define an unknown real-valued tolerance Z_i. In the regression setting, suppose that the tolerances Z_i follow the linear model

$$Z_i = x_i^{\mathrm{T}}\beta + \varepsilon_i, \tag{1}$$

where the errors ε_i are a random sample from a known symmetric cumulative distribution function F. If the observation y_i is generated according to the model

$$y_i = \begin{cases} 1 & \text{if } Z_i > 0, \\ 0 & \text{if } Z_i \leqslant 0, \end{cases} \tag{2}$$

then it is straightforward to check that $\mathrm{pr}\,(y_i = 1) = F(x_i^{\mathrm{T}}\beta)$. In the bioassay setting, Z_i can represent an insect's tolerance to a pesticide and the insect survives $(y_i = 1)$ if the tolerance exceeds some constant value.

In this context, it is natural to let the residual corresponding to y_i be defined as

$$\varepsilon_i(Z_i, \beta) = Z_i - x_i^{\mathrm{T}}\beta.$$

The properties of this 'latent data' residual are best understood in the setting of the probit model. In that case, it follows from (1) that the residuals $\varepsilon_1, \ldots, \varepsilon_N$ are, a priori, a random sample from a $N(0, 1)$ distribution, which provides a convenient base with which to compare the posterior distribution. To understand how the observations y_i change the distribution of these residuals, consider the posterior distribution of $\{\varepsilon_i(Z_i, \beta)\}$ conditional on β. From (2), the posterior density of ε_i is given by

$$\pi(\varepsilon_i | y, \beta) = \begin{cases} \dfrac{\phi(\varepsilon_i)}{\Phi(x_i^{\mathrm{T}}\beta)} I(\varepsilon_i > -x_i^{\mathrm{T}}\beta) & \text{if } y_i = 1, \\ \dfrac{\phi(\varepsilon_i)}{\Phi(-x_i^{\mathrm{T}}\beta)} I(\varepsilon_i < -x_i^{\mathrm{T}}\beta) & \text{if } y_i = 0, \end{cases} \tag{3}$$

where ϕ and Φ are the standard normal density and distribution function and $I(A)$ is the indicator function of the set A. This posterior is a truncated standard normal density, where the truncation point is the negative of the linear predictor $x_i^T\beta$. The shape of this density depends on the sign of the linear predictor and the value of the response y_i. The posterior density of ε_i will be significantly different from the prior density only when the observation is of the opposite sign to the linear predictor. In the case where $y_i = 1$, the posterior mean and variance of ε_i, conditional on β, are given by

$$E(\varepsilon_i | y_i = 1, \beta) = w_i, \quad \text{var}(\varepsilon_i | y_i = 1, \beta) = 1 - w_i(x_i^T\beta + w_i),$$

respectively, where $w_i = \phi(x_i^T\beta)/\Phi(x_i^T\beta)$. These moments will be significantly different from the prior moments, $E(\varepsilon_i) = 0$, var $(\varepsilon_i) = 1$, when w_i is large, or equivalently when the linear predictor $x_i^T\beta$ is smaller than some negative constant C. Also, when $y_i = 1$, the probability that the residual is larger than a pre-specified value $K > -x_i^T\beta$ is given by

$$\text{pr}(|\varepsilon_i| > K | y_i = 1, \beta) = \Phi(-K)/\Phi(x_i^T\beta).$$

These latent residuals are particularly interesting to study since they can be simulated and summarised as a by-product of the algorithm of Albert & Chib (1993). This Gibbs sampling algorithm relies on successive simulations from two conditional posterior distribution and is straightforward to implement. First, for a fixed value of β, the simulation of the components Z_i is made from independent truncated normal distributions. If $TN_{(a,b)}(\mu, \sigma^2)$ denotes the $N(\mu, \sigma^2)$ distribution truncated to the interval (a, b), then Z_i is simulated from $TN_{(0,\infty)}(x_i^T\beta, 1)$ if $y_i = 1$, and from $TN_{(-\infty,0)}(x_i^T\beta, 1)$ if $y_i = 0$. Secondly, for a given value of Z, the simulation of β, assuming a vague prior, is from

$$\mathcal{N}((X^TX)^{-1}X^TZ, (X^TX)^{-1}).$$

The sample of draws $\{(Z^{(g)}, \beta^{(g)})\}$ generated by this simulation can be used to obtain the posterior distribution of the latent residuals $\varepsilon_i(Z_i, \beta)$. Then, for a particular observation, the posterior distribution of the residual can be summarised by sample quantiles of the simulated values $\{\varepsilon_i(Z_i^{(g)}, \beta^{(g)})\}$ and the outlying probabilities of interest can be estimated by the fraction of simulated residuals that exceed a prespecified constant.

Alternative simulation-based estimates of the residual distributions can be obtained by use of the knowledge of the conditional distribution of the residuals given β. These are called Rao–Blackwellised estimates by Gelfand & Smith (1990). The marginal posterior density of the residual ε_i can be expressed as the mixture

$$\pi(\varepsilon_i | y) = \int \pi(\varepsilon_i | y, \beta)\pi(\beta | y)\, d\beta,$$

where $\pi(\varepsilon_i | y, \beta)$ is given in (3) and $\pi(\beta | y)$ is the marginal posterior density of the regression vector. It follows that a simulation estimate of the residual density is given by the sample mean of $\{\pi(\varepsilon_i | y, \beta^{(g)})\}$, where $\{\beta^{(g)}\}$ is a simulated sample from the Gibbs sampling run. Similarly, the distribution function of ε_i is given by

$$\Pi(\varepsilon_i | y) = \int \Pi(\varepsilon_i | y, \beta)\pi(\beta | y)\, d\beta,$$

where $\Pi(\varepsilon_i | y, \beta)$ is the distribution function corresponding to the density (3). From a simulated sample from β, one can compute a Rao–Blackwellised estimate of the distribution function of ε_i and this is useful in calculating quantiles or computing tail probabilities.

The latent residuals can also be used for other link models in which one cannot directly simulate the latent data. For example, consider the case of the logistic model where $\log\{p_i/(1-p_i)\} = x_i^{\mathrm{T}}\beta$ and a uniform prior is placed on β. Introduce latent data $\{Z_i\}$ from a logistic distribution with locations $\{x_i^{\mathrm{T}}\beta\}$ and scale parameter 1, where the observation y_i is again 1 or 0 if Z_i is positive or negative, respectively. The latent residuals $\{\varepsilon_i\} = \{Z_i - x_i^{\mathrm{T}}\beta\}$ are a priori a random sample from a standard logistic distribution. The marginal posterior density of ε_i, conditional on β, is given by (3), where the standard normal probability density function and cumulative distribution function are replaced by those of the standard logistic distribution. Unlike the probit case, it is not possible to simulate the values of the latent residuals directly. However, one can obtain a simulated sample from the posterior distribution of β using the Markov chain Monte Carlo algorithm of Dellaportas & Smith (1993). This simulated sample can be used to obtain the marginal densities and marginal distribution functions of the latent residuals. To illustrate, if $y_i = 1$, the posterior density of ε_i can be estimated by the sample mean of

$$\{f(\varepsilon_i)/F(x_i^{\mathrm{T}}\beta)\}I(\varepsilon_i > -x_i^{\mathrm{T}}\beta),$$

where $f(.)$ and $F(.)$ are the probability density function and cumulative distribution function of the standard logistic distribution and $\{\beta^{(g)}\}$ is a sample from the posterior distribution of β.

3. Graphical displays

To motivate the basic plots for the two types of residuals, consider the analysis of a small ($N = 20$) simulated data set. The data generating process is given by $y_i \sim \text{Bernoulli}(p_i)$, where

$$p_i = \text{pr}(y_i = 1) = \exp(\beta_0 + \beta_1 x_i)/\{1 + \exp(\beta_0 + \beta_1 x_i)\},$$

the covariate value x_i is uniformly distributed on $(-1, 1)$, and $\beta_0 = 0$, $\beta_1 = 3$. A logistic link function was chosen in this example to induce a few outliers in the data. In the analysis below, a probit link model was fitted with β assigned a diffuse prior. The following plots were based on a simulated sample of size 2000 from the posterior distribution of β.

Figures 1 and 2 plot the marginal posterior densities for the residuals $r_i = y_i - p_i$. The residual densities for the failures ($y_i = 0$) are graphed in Fig. 1. Figure 2 plots the entire set of 20 residual distributions using parallel boxplots. As in Fig. 1, the distributions are plotted against the fitted probabilities.

The basic pattern of Fig. 2 resembles the frequentist residual plot for binary data. The posterior medians of the residuals r_i fall on two parallel lines, where the top line corresponds to the observed successes and the bottom line to the failures. Note that the value of y_i determines the support of the posterior distribution of r_i; the support is $(0, 1)$ for $y_i = 1$ and $(-1, 0)$ for $y_i = 0$. Moreover, from observing Fig. 1, note that the shape of the residual distribution depends on its location. The distributions that are concentrated near the endpoints of the support region show substantive skewness. In particular, note that one of the residual densities corresponding to a failure ($y_i = 0$) has a mode at 0.

Outlying observations correspond to residual densities that have locations away from zero. A particular residual distribution for $y_i = 0$ that is a candidate for an outlier is labelled in Fig. 1. One way of gauging the relative sizes of these residuals is to compute the posterior probabilities that r_i exceeds in absolute value some positive constant K.

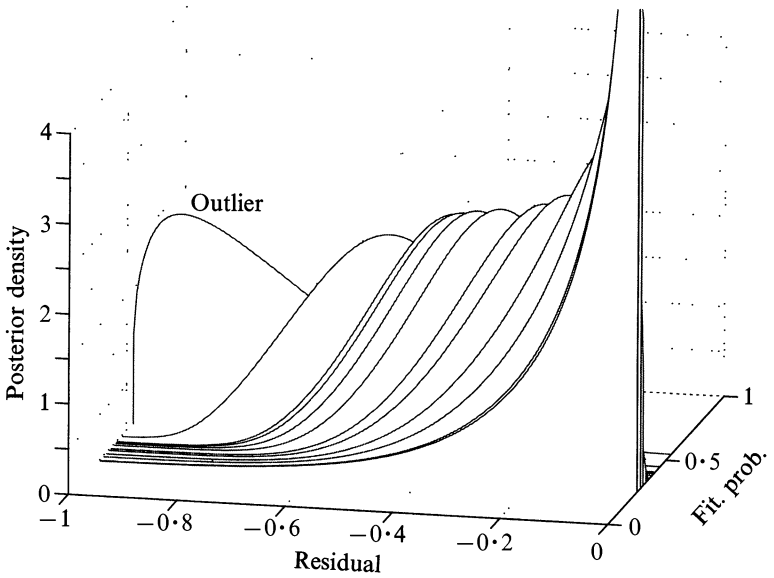

Fig. 1. Posterior densities for the residuals $r_i = y_i - p_i$ for simulated data set for observations where $y_i = 0$. The densities are plotted against the fitted probabilities $E(p_i|y)$.

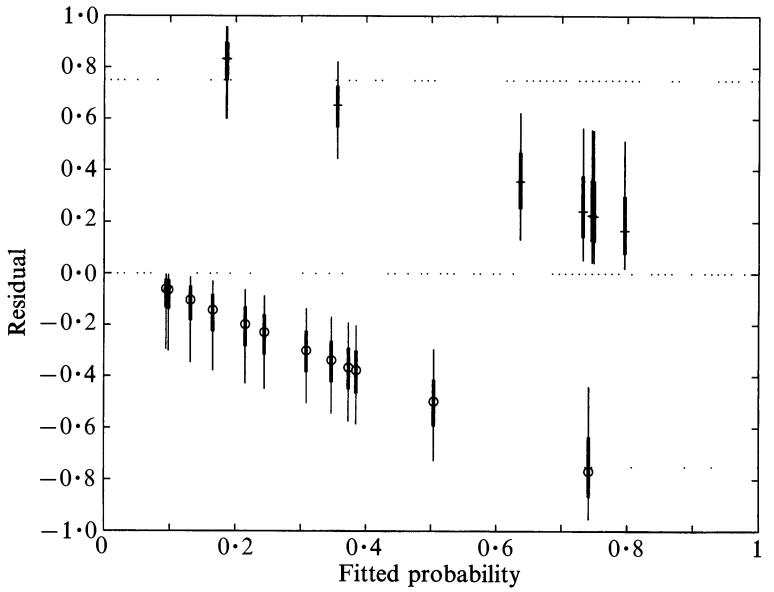

Fig. 2. Boxplots of posterior distributions for the residuals $r_i = y_i - p_i$ plotted agains the fitted probabilities $E(p_i|y)$ for simulated data set. The middle section of the boxplot corresponds to the quartiles, and the extreme values correspond to the 5th and 95th percentiles of the distribution. The boxplots drawn with a circle at the median correspond to the observations where $y_i = 0$, and the boxplots with a cross correspond to the observations where $y_i = 1$. Boxplots which significantly cross the dotted lines correspond to outliers.

Table 1. *Observations, fitted probabilities, and outlying probabilities for simulated data set*

Index	y_i	$E(p_i \mid y)$	pr $(\lvert r_i \rvert > 0{\cdot}75 \mid y)$	pr $(\lvert \varepsilon_i \rvert > 2 \mid y)$
1	0	0·098	0·000	0·028
2	0	0·504	0·032	0·050
3	1	0·733	0·002	0·033
4	1	0·186	0·752	0·193
5	0	0·309	0·000	0·034
6	1	0·637	0·006	0·038
7	0	0·347	0·000	0·036
8	0	0·166	0·000	0·028
9	1	0·187	0·747	0·190
10	0	0·244	0·000	0·031
11	0	0·132	0·000	0·028
12	1	0·356	0·190	0·072
13	0	0·742	0·541	0·159
14	1	0·750	0·002	0·032
15	0	0·095	0·000	0·029
16	1	0·746	0·002	0·033
17	0	0·373	0·001	0·038
18	0	0·385	0·001	0·039
19	0	0·215	0·000	0·030
20	1	0·796	0·002	0·031

These outlying probabilities are displayed in Table 1 for the value $K = 0{\cdot}75$. One can graphically see these outlying probabilities in Fig. 2. Parallel lines are drawn at residual values $-0{\cdot}75$ and 0·75. If the boxplot for a particular residual does not cross these lines, then the outlying probability is under 0·05. Residuals with large outlying probabilities correspond to boxplots that significantly cross these lines. From Table 1 and Fig. 2, we see that there are three 'success' observations, numbers 4, 9 and 13, with relatively large outlying probabilities.

The above posterior computations appear to be successful in identifying potential outliers. However, it can be difficult to interpret the sizes of the r_i distributions due to the bounded support. In particular, it is difficult to compare two residual distributions with different spreads. A more serious concern is that it is unclear how one should assign the value of the cut-off value K in the tail probability computation. The prior distribution of r_i is not known owing to the discrete nature of the response variable y. Thus it is difficult to understand what are reasonable sizes of the residuals before sampling.

Next consider the behaviour of the latent residuals for this data set. Figure 3 plots the posterior densities of the ε_i for the failures $(y_i = 0)$ against the posterior means of the probabilities p_i. Figure 4 summarises the collection of ε_i distributions using parallel boxplots. Since the marginal prior distribution of each residual is standard normal, these residual distributions are easier to interpret. Successes $(y_i = 1)$ with fitted probabilities close to one and failures $(y_i = 0)$ with fitted probabilities close to zero have residual distributions which resemble standard normal curves. In these cases, the observations have little influence on the distribution of the residuals. The posterior distribution of ε_i is significantly different from the $N(0, 1)$ prior distribution when the fitted probability is in conflict with the observation. One outlier candidate is labelled in Fig. 3. In Figs 3 and 4, this conflict is evident in the nonzero location, smaller standard deviation, and some skewness of the posterior residual distribution.

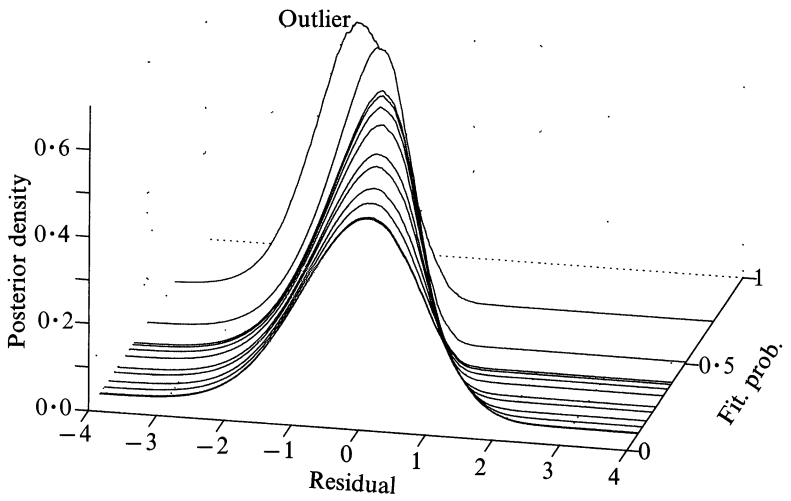

Fig. 3. Posterior densities for the residuals $\varepsilon_i = Z_i - x_i^T\beta$ plotted agains the fitted probabilities $E(p_i|y)$ for simulated data set for observations where $y_i = 0$.

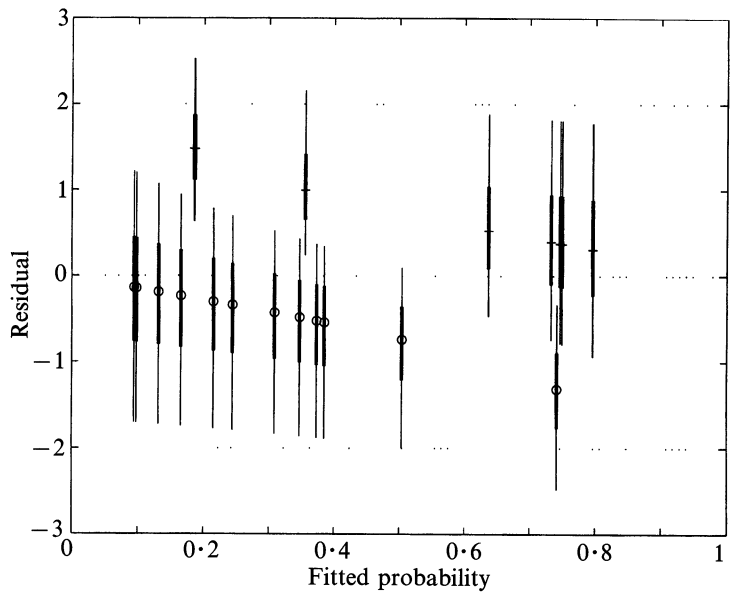

Fig. 4. Boxplots of posterior distributions for the residuals $\varepsilon_i = Z_i - x_i^T\beta$ plotted against the fitted probabilities $E(p_i|y)$ for simulated data set. The boxplots drawn with a circle at the median correspond to the observations where $y_i = 0$, and the boxplots with a cross correspond to the observations where $y_i = 1$. Boxplots which cross the dotted lines at -2 and 2 correspond to outliers.

One clear way of seeing this conflict is by the computation of the outlying probabilities pr $(|\varepsilon_i| > K|y)$. These probabilities are given in Table 1 for $K = 2$. As in Fig. 2, one can see by inspection of Fig. 4 whether particular observations are outlying. If a partic-

ular boxplot crosses one of the parallel lines at $\varepsilon = -K$ and $\varepsilon = K$, then this particular observation deserves special attention.

There are some general observations that can be made about the ε distributions in this regression setting. First, when the fitted probability is in strong agreement with the observation, the residual distribution is essentially a standard normal distribution with only one tail truncated; the effect of this truncation is to lower the outlying probability to half its prior value. When the fitted probability is 0·5, the linear predictor $x_i^T \beta$ is approximately 0 and the residual distribution is truncated at 0. The effect of this particular truncation is to keep the outlying probability at its prior value. Lastly, when there is significant conflict between y_i and the fitted probability, there is severe truncation in the residual distribution, causing a large outlying probability. Thus this method will set apart particular observations whose fitted probabilities are close to 0 or 1 where the binary response is in conflict.

4. Real data examples

4·1. *Binary regression*

Now the use of Bayesian residual distributions is illustrated in the context of real data. Consider the study of Brown (1980) in which each of 53 prostatic cancer patients had a laparotomy to see if the cancer had spread to the surrounding lymph nodes. The binary response measures the presence or absence of nodal involvement and one is interested in exploring the relationship between nodal involvement and five other variables. The possible covariates include the age of the patient, the level of serum acid phosphates, the result of an X-ray examination, the size of the tumour determined by a rectal examination, and a summary of the pathological grade of the tumour. A detailed analysis of this data set from a frequentist perspective has been presented by Collett (1991).

Consider two models to these data. Model 1, which will be referred to as the badly fitting model, includes two of the poorer predictor variables, age and grade, and gives a deviance of 65·26 on 50 degrees of freedom. In contrast, Model 2 is a relatively well fitting model that includes main effects for the four covariates log (acid), X-ray, size and grade, with a deviance of 47·52 on 48 degrees of freedom. For Model 1, Fig. 5 graphs the residual distributions for the ε_i residual definition. As in the simulated data example, the cut-off value 2 is used to distinguish outliers.

Certain features of this plot are noteworthy. First, the range of fitted probabilities is relatively small, which is an indication of the poor predictive performance of the model. All of the observed successes and observed failures have similar fitted probabilities. Secondly, there is little overlap of the boxplots with the outlier values of -2 and 2. The posterior probability that ε_i exceeds 2 in absolute value is under 0·15 for all observations. Since these posterior probabilities are close to the prior probabilities of 0·05, the conclusion is that no observations are unusual in this badly fitting model.

Figure 6 presents residual distributions from Model 2. This model appears to fit the data better since, for many observations, the fitted probabilities are close to the response values of 0 and 1. When the fitted probabilities are near the extreme values as in this example, one is more likely to observe large residuals. Indeed, from inspection of the boxplots, there are three observed successes and one observed failure with large residuals that overlap the outlier bounds. The responses, fitted probabilities, and outlying probabilities for these four observations are given in Table 2. All four observations have tail probabilities that significantly exceed the prior outlying probability of 0·05.

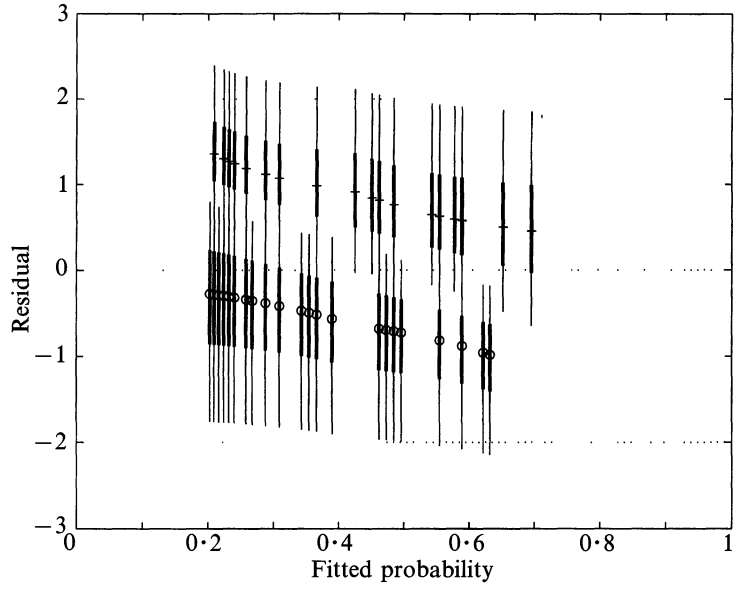

Fig. 5. Posterior boxplots for $\varepsilon_i = Z_i - x_i^T\beta$ residual distributions for badly-fitting model.

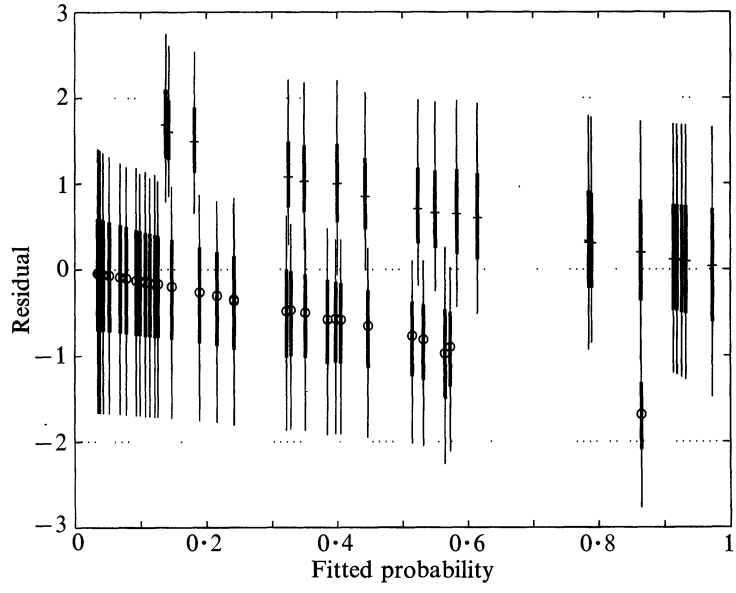

Fig. 6. Posterior boxplots for $\varepsilon_i = Z_i - x_i^T\beta$ residual distributions for well-fitting model.

Table 2 also provides some classical outlier statistics for these four observations. The values of the deviance residual, defined in § 1, are given. If one plots the entire set of deviance residuals, these four values (2·02, 1·96, 1·87, −2·03) do not stand out. As mentioned in § 1, it is difficult to interpret the sizes of these residuals, since the sampling distributions are unknown. Alternative classical methods can set apart these four observations. For example, one method is to compute the change in deviance if each observation

Table 2. *Observations, fitted probabilities, outlying probabilities, deviance residuals, and change in deviance for four observations in cancer data set*

Index	y_i	$E(p_i \mid y)$	pr $(\lvert\varepsilon_i\rvert > 2 \mid y)$	d_i	ΔD
9	1	0·138	0·304	2·02	4·92
26	1	0·143	0·234	1·96	4·24
35	1	0·182	0·197	1·87	3·98
37	0	0·865	0·299	−2·03	4·97

in turn is deleted from the data set. The changes in deviance given in Table 2 for these four observations are in the range 4 to 5 and these are very large compared to the remaining observations. However, this information is not provided by inspection of the set of deviance or Pearson residuals.

4·2. *A longitudinal random effects model*

The residual definitions developed in this paper can be generalised in a straightforward manner to longitudinal random effects models with a binary response. In this section, the use of the latent data residuals is illustrated in the analysis of data from a crossover trial (Kenward & Jones, 1987). In this study, 86 patients were treated for the relief of pain in primary dysmennorrhoea. Each patient receives each of three treatments during 3 periods and, at the end of each time period, the treatment is rated as either giving no relief or some relief. Let y_{it} denote the response and p_{it} the probability that subject i experiences some relief during period t. Then one probit model is given by

$$\text{pr}\,(y_{it} = 1 \mid b_i) = \Phi(x_i^{\mathrm{T}}\beta + b_i) \quad (t \leqslant 3, i \leqslant 86), \tag{4}$$

where β is a regression vector which models possible effects due to treatment, period, and crossover effects, x_{it} is the corresponding covariate vector, and b_i is the subject-specific tolerance to the drugs. The 86 random effects $\{b_i\}$ are assumed to be a random sample from a normal distribution with mean 0 and unknown variance σ^2. The model is completed by assigning vague prior distributions to the regression vector β and the random effects variance σ^2.

To estimate the model in (4), latent data $\{Z_{it}\}$ are introduced, with $Z_{it} \mid b_i \sim N(x_{it}^{\mathrm{T}}\beta + b_i, 1)$, representing the patients' relief from the ailment measured on a continuous scale. We have $y_{it} = 1$ if $Z_{it} > 0$, and $y_{it} = 0$ otherwise. In this setting, the latent data residual can be defined as $\varepsilon_{it}(Z_{it}, \beta, \sigma^2) = Z_{it} - x_{it}^{\mathrm{T}}\beta$. Conditional on σ^2, these residuals have independent $\mathcal{N}(0, 1 + \sigma^2)$ prior distributions. It is possible to proceed as described earlier and, for example, compute the outlying probability,

$$\text{pr}\,\{\lvert\varepsilon_{it}\rvert > K(1+\sigma)^{\frac{1}{2}} \mid y\}, \tag{5}$$

to assess whether a particular observation is large. Using a slight generalisation of the Markov chain Monte Carlo algorithm described in § 2·2, (5) is obtained using the simulated sample from the posterior distribution of all unobservables.

Albert & Chib (1995), in their analysis of these data, found that there were significant treatment effects and that the period and carryover effects were nonsignificant. Specifically, there was a significant increase in relief from level one of the treatment (placebo) to level two (low dose of the drug), and a modest increase in relief from level two (low dose) to

level three (high dose) of the treatment. In addition, the sizes of the random effects were significant. There exists variation in the responses due to treatment effects and due to differences between patients.

To check the fit of this model, one can look at the set of Bayesian residuals $\{\varepsilon_{it}\}$. Table 3 gives the probability that $|\varepsilon_{it}| > 2$ for some unusual cases. Each line of the table gives the observed binary response for treatments 1, 2, 3, the posterior mean of the probability of response for all three treatments, and the corresponding outlying probabilities. A typical response for these data for the three treatments is (0, 1, 1) and observed responses that deviate from this typical response can show large outlying probabilities. As in the previous example, the prior probability that each residual is outlying is 0·05. The large tail probabilities given in this table indicate that the responses are not adequately fitted by the random effects model. One may next wish to consider a more complicated model by adding new covariates or a higher-dimensional random effects structure.

Table 3. *Binary observations, fitted probabilities, and outlying probabilities for some unusual cases in cross-over data set*

Observation			Fitted probabilities			Outlying probabilities		
Treatment			Treatment			Treatment		
1	2	3	1	2	3	1	2	3
0	0	0	0·12	0·45	0·67	0·06	0·10	0·16
1	0	0	0·22	0·60	0·78	0·04	0·08	0·15
0	1	0	0·22	0·61	0·78	0·04	0·02	0·15
1	1	1	0·32	0·90	0·87	0·15	0·05	0·05

5. CONCLUDING REMARKS

This paper has discussed the use of two Bayesian residuals in the detection of outlying observations in binary response regression models. The posterior distributions $\{r_i\}$, plotted against the predicted probabilities, provide a simple graphical approach to detecting extreme observations. Similar graphical information can be gleaned from the posterior distribution of the latent residual ε_i. In addition, in the probit case, owing to the fact that the prior distribution of the latent residual is standard normal, it is simple to compare one's prior and posterior beliefs about outlying data, especially with respect to the occurrence of tail events.

Other schemes could be designed. For example, we could classify outliers according to, say, the Kullback divergence (Kullback, 1959) between the prior and posterior residual distribution for a particular observation. Finally, the approach described here can be easily extended to investigate whether particular subsets of observations are inconsistent with the regression model. This could be done by computing the posterior probability of the intersection of the events $|\varepsilon_i| > K$ for a set of subscripts i. These probabilities could be helpful in the detection of outliers that are masked by other observations.

REFERENCES

ALBERT, J. & CHIB, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Am. Statist. Assoc.* **88**, 669–79.

ALBERT, J. & CHIB, S. (1995). Bayesian probit modeling of binary repeated measures data with an application to a cross-over trial. In *Bayesian Biostatistics*, Ed. D. A. Berry and D. K. Stangl. To appear. New York: Marcel Dekker.

ATKINSON, A. C. (1981). Two graphical displays for outlying and influential observations in regression. *Biometrika* **68**, 357–63.

BROWN, B. W. (1980). Prediction analysis for binary data. In *Biostatistics Casebook*, Ed. R. J. Miller, B. Efron, B. W. Brown and L. E. Moses, pp. 3–18. New York: Wiley.

CHALONER, K. (1991). Bayesian residual analysis in the presence of censoring. *Biometrika* **78**, 637–44.

CHALONER, K. & BRANT, R. (1988). A Bayesian approach to outlier detection and residual analysis. *Biometrika* **75**, 651–9.

COLLETT, D. (1991). *Modelling Binary Data*. London: Chapman and Hall.

DAVISON, A. C. (1988). Contributions to the discussion of a paper by J. B. Copas. *J. R. Statist. Soc.* B **50**, 258–9.

DELLAPORTAS, P. & SMITH, A. F. M. (1992). Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling. *Appl. Statist.* **42**, 443–59.

DUFFY, D. (1990). On continuity-corrected residuals in logistic regression. *Biometrika* **77**, 287–93.

FOWLKES, E. B. (1987). Some diagnostics for binary logistic regression with smoothing. *Biometrika* **74**, 503–15.

GELFAND, A. E. & SMITH, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Am. Statist. Assoc.* **85**, 398–409.

KENWARD, M. G. & JONES, B. (1987). Modelling binary data from a theoretical crossover trial. *Statist. Med.* **10**, 1607–19.

KULLBACK, S. (1959). *Information Theory and Statistics*. New York: Wiley.

PREGIBON, D. (1981). Logistic regression diagnostics. *Ann. Statist.* **9**, 705–24.

ZELLNER, A. (1975). Bayesian analysis of regression error terms. *J. Am. Statist. Assoc.* **70**, 138–44.

ZELLNER, A. & MOULTON, B. R. (1985). Bayesian regression diagnostics with applications to international consumption and income data. *J. Economet.* **29**, 187–211.

[*Received March* 1994. *Revised November* 1995]