Statistical Practice

Bayesian Binomial Regression: Predicting Survival at a Trauma Center

Edward J. BEDRICK, Ronald CHRISTENSEN, and Wesley JOHNSON

Standard methods for analyzing binomial regression data rely on asymptotic inferences. Bayesian methods can be performed using simple computations, and they apply for any sample size. We provide a relatively complete discussion of Bayesian inferences for binomial regression with emphasis on inferences for the probability of "success." Furthermore, we illustrate diagnostic tools, perform model selection among nonnested models, and examine the sensitivity of the Bayesian methods.

KEY WORDS: Bayesian analysis; Importance sampling; Kullback–Leibler divergence; Logistic regression; Model selection; Prediction; Probit analysis; Regression diagnostics.

1. INTRODUCTION

The recent development of Monte Carlo methods has eliminated most of the difficulties historically associated with Bayesian analyses of nonlinear models. This paper illustrates the simplicity of a fully Bayesian approach to binomial regression models using data from the University of New Mexico Trauma Center. In particular, we discuss a prior specification that focuses on eliciting binomial probabilities, rather than specifications for the more esoteric regression coefficients. We use simple Monte Carlo methods for prediction, inferences on regression coefficients and probabilities, diagnostics, link selection, and sensitivity analysis of the prior. A complete analysis can be handled easily and accurately within this framework.

Most of the methods discussed have appeared elsewhere. Leonard (1972) discussed Bayesian hierarchical models for binomial data. Zellner and Rossi (1984) gave an overview of Bayesian methods for binomial regression models. Johnson (1985) introduced predictive case deletion diagnostics for binomial regression. We integrate their ideas along with Bedrick, Christensen, and Johnson's (1996) (hereafter referred to as BCJ) ideas on specifying priors to provide a variety of tools appropriate for analyzing binomial response data.

Consider regression data $(y_i, x_i'), i = 1, \ldots, n$, where the y_is are success proportions from independent binomial N_i random variables and the x_is are known k vectors of covariates. The probability of success for any single trial y with covariate x is $r(x'\beta)$, that is, $r(x'\beta) \equiv p(y = 1|x, \beta) = p$, where β is an unknown k vector of regression coefficients. The function $r(\cdot)$ can be an arbitrary cdf, but we will assume without much loss of generality that $r(\cdot)$ corresponds to either the logistic, probit, or complementary log–log models, that is,

$$r(x'\beta) = \begin{cases} e^{x'\beta}/[1 + e^{x'\beta}] & \text{logistic} \\ \Phi(x'\beta) & \text{probit} \\ 1 - \exp[-e^{x'\beta}] & \text{complementary log}-\log. \end{cases}$$

The link function is $r^{-1}(p) = x'\beta$, where $r^{-1}(p) = \log\{p/(1-p)\}, \Phi^{-1}(p)$, and $\log\{-\log(1-p)\}$ for the three models, respectively. The likelihood for data $Y = (y_1, \ldots, y_n)'$ is

$$L(\beta|Y) \equiv \prod_{i=1}^{n} L(\beta|y_i)$$

$$\equiv \prod_{i=1}^{n} \binom{N_i}{N_i y_i} [r(x_i'\beta)]^{N_i y_i} [1 - r(x_i'\beta)]^{N_i - N_i y_i}$$

For a prior $\pi(\beta)$ the posterior of β is

$$\pi(\beta|Y) = \frac{L(\beta|Y)\pi(\beta)}{\int L(\beta|Y)\pi(\beta)\, d\beta}.$$

Most interesting aspects of a Bayesian analysis are obtained from various integrals involving $\pi(\beta|Y)$. Such integrals are intractable, so we use simulation methods to obtain approximations. Two popular approaches use importance sampling and Gibbs sampling (Zellner and Rossi 1984; Dellaportas and Smith 1993). Alternative approaches based on Laplace approximations (Tierney and Kadane 1986) and numerical integration (Smith, Skene, Shaw, Naylor, and Dransfield 1985) are available, but are less commonly used.

Simulation methods yield a discrete approximation to the posterior distribution taking values β^i with probability $\tilde{q}_i$, $i = 1, \ldots, t$ as discussed in Section 3. Given a function $h(\beta)$ the posterior expectation $\theta_h \equiv \mathrm{E}\{h(\beta)|Y\}$ is approximated

Edward J. Bedrick is Associate Professor and Ronald Christensen is Professor, Department of Statistics, University of New Mexico, Albuquerque, NM 87131 (E-mail: fletcher@math.unm.edu). Wesley Johnson is Professor, Division of Statistics, University of California at Davis, Davis, CA 95616. The authors thank Turner Osler for the use of his data and for donating his time to the prior elicitation process. They also thank the associate editor and the referees for their valuable comments.

by

$$\bar{\theta}_h = \sum_{i=1}^{t} \tilde{q}_i h(\beta^i). \tag{1.1}$$

Typically, the Strong Law of Large Numbers applies to give almost sure convergence as t increases.

Section 2 discusses standard Bayesian inference, with emphasis on the predictive distribution. It includes influence measures and a procedure for selection of the appropriate link function. Section 3 discusses computational issues. Section 4 contains concluding remarks and other suggested source material.

2. BAYESIAN INFERENCE

2.1 Specifying the Prior

Several methods of specifying priors for binomial regression problems have been proposed. The standard approach has been to assume either a normal prior, or the diffuse prior $\pi(\beta) = 1$ for the regression coefficients. These are convenient in large sample situations where the posterior on β is approximately normal. See Zellner and Rossi (1984) for relevant discussion.

Another form of prior focuses on the assessment of "success" probabilities for various choices of covariate values, rather than on the assessment of regression coefficients. For binomial regression problems with a single predictor variable Tsutakawa (1975), Tsutakawa and Lin (1986), and Grieve (1988) consider joint prior distributions on two probabilities of success. In their examples independent beta distributions are placed on the two probabilities, but in addition, they consider that one of these probabilities may dominate the other, thus restricting the parameter space. The distribution of β is induced from this distribution on probabilities using the change-of-variables technique.

We follow BCJ's extension of Tsutakawa's approach to generalized linear models (GLMs) with multiple covariates. A proper prior for the binomial response model $r(x'\beta) = p$ with k regression parameters is obtained from a prior elicited on success probabilities $\tilde{p}_i$ at k suitably selected covariate vectors $\tilde{x}_i$. The $\tilde{x}_i$s should be chosen in the anticipated range of the observed predictor variables, and should also be amenable to expert opinion. Let $\tilde{X}$ be the $p \times p$ nonsingular matrix with $\tilde{x}_i'$ in the ith row, $i = 1, \ldots, p$. Suppose that the elicited prior on $\tilde{p} = (\tilde{p}_1, \ldots, \tilde{p}_k)'$ is given by $\pi_0(\tilde{p})$, and define the 1–1 transformation of β by $\tilde{p} = R(\tilde{X}\beta)$, where R is the vector transformation that applies r to each element. Using the change-of-variables technique the induced prior on β is

$$\pi(\beta) = \pi_0(R(\tilde{X}\beta)) \left| \frac{\partial R(\tilde{X}\beta)}{\partial \beta} \right|. \tag{2.1}$$

BCJ call such priors conditional means priors (CMPs), and they elaborate on their use in considerable detail.

We agree with Tsutakawa and Lin (1986) that eliciting information about success probabilities should be much easier than eliciting information about regression coefficients. This is especially true if one entertains the possibility of two or more models such as logistic regression versus probit regression. The regression coefficients for these two models require separate elicitations because the regression coefficients have different meanings, whereas if one has elicited a prior for the probabilities, it is straightforward to induce the requisite prior on β for either model.

A key feature in the Tsutakawa (1975) approach as generalized by BCJ is assuming prior independence of the $\tilde{p}_i$s. This assumption might be unreasonable if the $\tilde{x}_i$ are "too close" together (Grieve 1988). There are also technical difficulties if they are "too far apart." BCJ (Section 5) examined these issues in detail, so we will not dwell on them here.

Beta distributions are natural priors for probabilities, but other priors could be considered. In our examples we specify values of $\tilde{N}_i > 0$ and $0 < \tilde{y}_i < 1$, and place independent Beta$(\tilde{N}_i\tilde{y}_i, \tilde{N}_i - \tilde{N}_i\tilde{y}_i)$ priors on the $\tilde{p}_i$, regardless of the link function. Given (2.1) the induced prior on β has the form

$$\pi(\beta) \propto \prod_{i=1}^{k} [r(\tilde{x}_i'\beta)]^{\tilde{N}_i\tilde{y}_i - 1} [1 - r(\tilde{x}_i'\beta)]^{\tilde{N}_i - \tilde{N}_i\tilde{y}_i - 1} \dot{r}(\tilde{x}_i'\beta)$$

where $\dot{r}(\cdot)$ is the first derivative of the function $r(\cdot)$. In the case of logistic regression

$$\pi(\beta) \propto \prod_{i=1}^{k} [r(\tilde{x}_i'\beta)]^{\tilde{N}_i\tilde{y}_i} [1 - r(\tilde{x}_i'\beta)]^{\tilde{N}_i - \tilde{N}_i\tilde{y}_i} \tag{2.2}$$

which has the same form as the likelihood function. The prior (2.2) is a data augmentation prior (DAP), so named because the likelihood times the prior has the form of a likelihood with additional "prior" data $(\tilde{y}_i, \tilde{N}_i), i = 1, \ldots, k$. So for the logistic model we can think of the parameters of each Beta distribution as a prior sample size $\tilde{N}_i$ and a prior proportion of successes $\tilde{y}_i$ corresponding to the vector of predictors $\tilde{x}_i$.

To motivate Tsutakawa's and BCJ's approach, consider a two-sample proportion problem. Imagine that the attending surgeon at a trauma center sees patients from two populations: gunshot wounds $(i = 1)$ and car crashes $(i = 2)$. Let $\tilde{p}_i$ be the probability of surviving one's wounds in population i, that is $\tilde{p}_1 = \Pr(\text{Survive}|\text{Gunshot})$ and $\tilde{p}_2 = \Pr(\text{Survive}|\text{Crash})$, and set $\tilde{p}_i = r(\tilde{x}_i'\beta)$, $i = 1, 2$ where $\tilde{x}_1' = (1, 0)$, $\tilde{x}_2' = (1, 1)$, and $\beta = (\beta_1, \beta_2)'$. Note that the model places no restriction on the probabilities, so the link is irrelevant.

The surgeon specifies that $\tilde{p}_1 \sim$ Beta(4, 4) and $\tilde{p}_2 \sim$ Beta(4, 1), reflecting (in part) her beliefs that about 80% $= \mathrm{E}(\tilde{p}_2) = 4/(4+1)$ of the crash victims and 50% $= \mathrm{E}(\tilde{p}_1)$ of the gunshot victims survive. Our prior assumes that $\tilde{p}_1$ and $\tilde{p}_2$ are independent. Standard textbooks routinely make this assumption (Lee 1989, p. 162). Thus, if we were told the value of $\tilde{p}_1$, we would not be inclined to revise our thinking about $\tilde{p}_2$. That seems reasonable if we are told that $\tilde{p}_1$ is near its expected value .5. It seems less reasonable if we are told, say, that $\tilde{p}_1 \geq .95$. Knowing that $\tilde{p}_1 \geq .95$ would probably make us want to revise our distribution of $\tilde{p}_2$ to make larger values more probable. However, .95 is 2.7 prior standard deviations above the prior mean for $\tilde{p}_1$, so this event

is unlikely under the prior specification. If $\tilde{p}_1 \geq .95$ is more likely than the original prior specification allows, the entire prior should be recalibrated, at which point the independence assumption may be called into question. However, if those situations that might cause concern about the independence assumption are thought unlikely, then we believe that the independence assumption is reasonable. Lack of independence could also occur if gunshot victims are thought to be medically similar to crash victims. In this case a prior might require a distribution for $\tilde{p}_1 - \tilde{p}_2$ that is concentrated near 0, and our prior is not appropriate.

Regardless of whether the independence assumption is made, the distribution of β is easily induced from the distribution of $(\tilde{p}_1, \tilde{p}_2)'$ using (2.1). Although any link can be used, the prior on β depends on the chosen link. This reinforces our view that one should elicit prior information about the only easily interpretable quantities here, the survival probabilities.

2.2 Trauma Data

We analyzed data on a random sample of 300 patients admitted to the University of New Mexico Trauma Center between the years 1991 and 1994. For each patient we have their injury severity score (ISS), their revised trauma score (RTS), their age (AGE), the predominant type of injury (TI), that is, whether it was blunt (TI = 0), for example, the result of a car crash, or penetrating (TI = 1), for example, gunshot wounds, and whether the patient eventually survived. The ISS (Baker et al. 1974) is an overall index of a patient's injuries based on the approximately 1,300 injuries cataloged in the Abbreviated Injury Scale. The ISS can take on values from 0 for a patient with no injuries to 75 for a patient with severe injuries in three or more body areas. The RTS is an index of physiologic injury, and is constructed as a weighted average of an incoming patient's systolic blood pressure, respiratory rate, and Glasgow Coma Scale (Champion et al. 1989). The RTS takes on values from 0 for a patient with no vital signs to 7.84 for a patient with normal vital signs.

Figure 1 gives side-by-side boxplots comparing the 278 survivors and 22 fatalities on RTS, ISS, and AGE. Seventeen of the 225 patients with blunt injuries died. Five of the 75 patients with penetrating injuries died.

These data were provided by Dr. Turner Osler, a trauma surgeon at the University of Vermont and former member of the University of New Mexico Trauma Center. We use

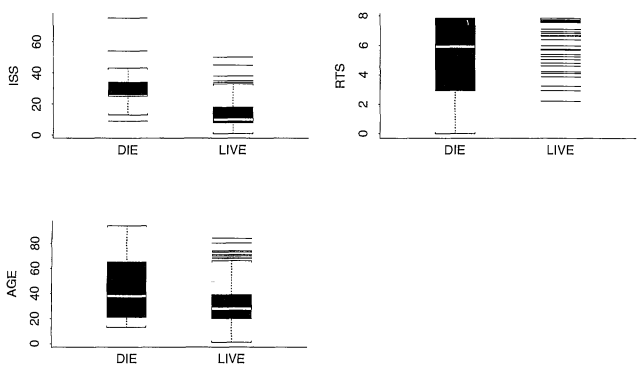

Figure 1. Side-by-Side Boxplots for the Trauma Data.

Table 1. Prior Specification

	Design for prior						Beta (a_i,b_i)	
i	$\tilde{x}_i'$						a_i	b_i
1	1	25	7.84	60	0	0	1.1	8.5
2	1	25	3.34	10	0	0	3.0	11.0
3	1	41	3.34	60	1	60	5.9	1.7
4	1	41	7.84	10	1	10	1.3	12.0
5	1	33	5.74	35	0	0	1.1	4.9
6	1	33	5.74	35	1	35	1.5	5.5

a logistic regression model proposed by Dr. Osler to estimate the probability of a patient's death using an intercept, predictors ISS, RTS, AGE, and TI, along with an interaction between AGE and TI. Henceforth $r(\cdot)$ indicates the logistic transformation. Similar logistic models are used by trauma centers throughout the United States (Champion, Sacco, and Copes 1991; Osler 1993). Dr. Osler's expert opinions formed the basis for our prior.

To induce a proper prior distribution on the six-dimensional vector β we require a joint distribution on death probabilities for six sets of conditions: $\tilde{x}_i' = (1, \text{ISS}_i, \text{RTS}_i, \text{AGE}_i, \text{TI}_i, \text{AGE}_i \times \text{TI}_i)$. Based on discussions with our expert and plots of the data we defined a 2^4 factorial having ISS at levels 25 and 41, RTS at levels 3.34 and 7.84, AGE at levels 10 and 60, and TI at levels 0 and 1. The idea was to pick values of the variables that were relatively extreme within the data, but still had substantial probabilities for both death and survival. The prior conditions were chosen as a 1/4 replicate of this 2^4 with two center points. However, the center points were taken to be values that could exist. In particular, TI is a binary variable, so one "center point" was taken with TI = 0 and the other with TI = 1. The prior design is given in Table 1. Beta priors were found to be suitable for the $\tilde{p}_i$s with parameters given in Table 1.

Figure 2 gives plots of the priors and posteriors on the $\tilde{p}_i$s. Relative to the amount of data, the priors are neither weak nor overwhelming, but generally the priors were consistent with the data. Certainly, the posteriors are not describing behavior that is inconsistent with prior beliefs. The posterior density for $\tilde{p}_i$ was obtained by smoothing a random sample from the discrete distribution that takes on the value $r(\tilde{x}_i'\beta^j)$ with probability $\tilde{q}_j$.

Our initial discussion with Dr. Osler involved eliciting 1st, 50th, and 99th percentiles for each $\tilde{p}_i$. These overspecify a beta distribution. We wrote a computer program to find the beta distribution that most nearly satisfied the specifications, plotted these distributions, and validated them with our expert.

The first probability $\tilde{p}_1$ corresponds to an individual that "has good physiology, is 'not badly hurt,' does not have a lot of reserve," and for whom there is "added uncertainty due to age." The Beta(1.1, 8.5) suitably reflects Dr. Osler's uncertainty about $\tilde{p}_1$. The median of his prior is .09. The second type of individual "has bad physiology, is very ill, but is young and resilient, and is not so badly hurt." The prior for this probability is Beta(3, 11) with median .20. Incidentally, "bad physiology" and "very ill' apparently refer

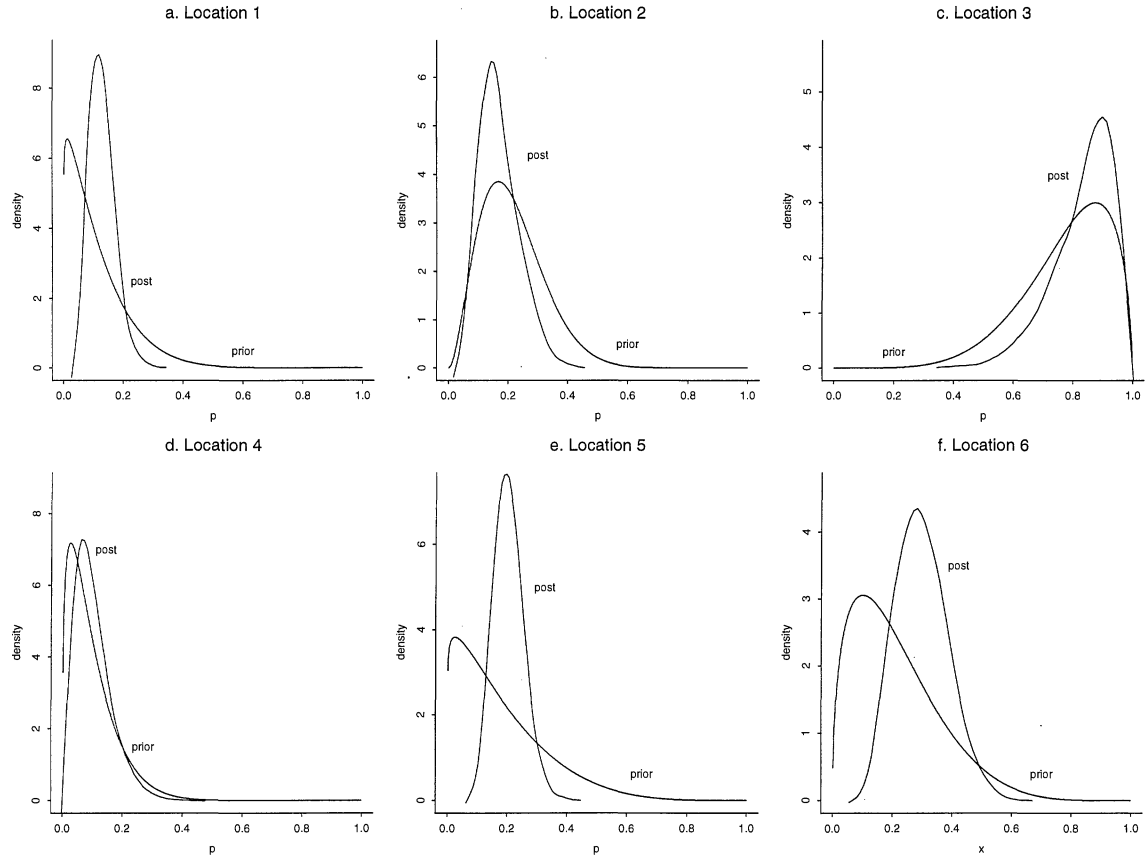

Figure 2. Prior and Posterior (Post) Distributions at Six Covariate Combinations.

to low RTS scores, while how badly hurt one is relates to ISS. The third individual has "bad physiology, a pretty bad injury, and there is much more uncertainty here due to the age factor." The prior is Beta(5.9, 1.7) with a median of .80. Individual 4 "is young, resilient, and has a big injury." The prior is Beta(1.3, 12) with a median of .07.

Dr. Osler had more difficulty with the fifth and sixth types of individuals because their conditions were both less extreme and more related than those already considered. The priors are Beta(1.1, 4.9) with median .15, and Beta(1.5, 5.5) with a median of .19, respectively.

The assumption of independence of the $\tilde{p}_i$s seems reasonable, with the possible exception of $\tilde{p}_5$ and $\tilde{p}_6$. If our expert were told that $\tilde{p}_5 = .3$, he would want to revise his probability for $\tilde{p}_6$ upwards. This is because he is confident that the difference between these two probabilities, $\tilde{p}_6 - \tilde{p}_5$, is positive but reasonably small, while he is less certain about the magnitude of the probabilities themselves. Having $\tilde{p}_6 - \tilde{p}_5$ small but positive reflects his perception that penetrating injuries are worse than blunt ones, but not a lot worse.

Because of our concern about possible lack of independence for the two values $\tilde{p}_6$ and $\tilde{p}_5$ only, we also considered priors in which the information about $\tilde{p}_6$ or, alternatively, $\tilde{p}_5$ was left out of the specification. This results in a partial prior (see BCJ, Section 4.1) that is an improper DAP using five prior observations instead of the six required for a proper DAP. We found that all inferences were essentially the same for the priors, so we have presented results only for the full prior. See Section 2.7 for further discussion of this issue.

2.3 Predictive Probabilities

The predictive probability of success in a new trial y with covariate x is

$$p(y = 1|Y, x) = \mathrm{E}[r(x'\beta)|Y, x] = \int r(x'\beta)\pi(\beta|Y)\, d\beta. \qquad (2.3)$$

Using the discrete approximation to the posterior gives

$$p(y = 1|Y, x) \doteq \sum_{j=1}^{t} r(x'\beta^j)\tilde{q}_j.$$

Figure 3 presents predictive probabilities of death as a function of ISS for blunt and penetrating traumas. These are given for four combinations of RTS and age. For 60 year olds there is essentially no difference in the probability of death due to blunt or penetrating injury. However, for 10 year olds the probability of death is higher for a penetrating injury. Note also, for example, that in Figure 3b the value plotted for ISS = 25, and Blunt is the mean for the posterior density in Figure 2b because the conditions of $\tilde{x}_2$ correspond to those conditions in Figure 3b.

The predictive probability of "success" can be interpreted in two ways. It is the subjective probability of death at a

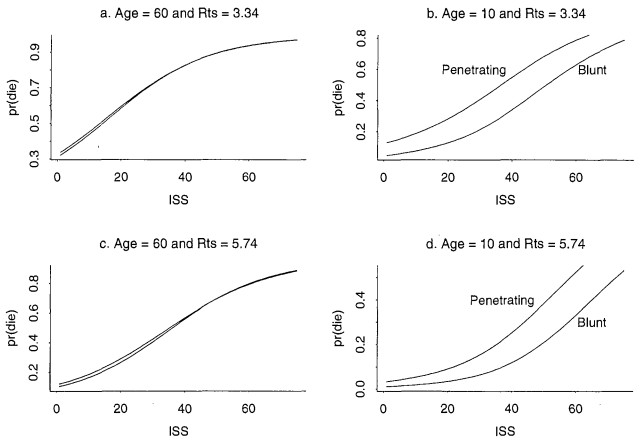

Figure 3. Predictive Distributions of Dying for Four Combinations of Age and RTS.

given set of covariates. It is also the Bayes estimate of the proportion of deaths among those with the given covariate values. With the second interpretation one may be interested in interval estimates. Geisser (1982) established that posterior interval estimates for probabilities can be viewed as (asymptotic) prediction intervals for the proportion of successes from a large number of future trials.

2.4 Inference for Regression Coefficients

The posterior mean $\mathrm{E}(\beta|Y)$ and covariance matrix

$$\mathrm{cov}(\beta|Y) = \mathrm{E}(\beta\beta'|Y) - \mathrm{E}(\beta|Y)\mathrm{E}(\beta|Y)'$$

of the regression coefficients are approximated by $\hat{\beta} = \sum_{j=1}^{t} \beta^j \tilde{q}_j$ and

$$\widehat{\mathrm{cov}}(\beta|Y) = \left[\sum_{j=1}^{t} \beta^j \beta^{j\prime} \tilde{q}_j\right] - \hat{\beta}\hat{\beta}',$$

respectively. The cdf of any β_h and histograms for approximating its marginal posterior density can be obtained from probabilities of the form

$$\begin{aligned}\Pr(a < \beta_h \le b|Y) &= \int I(a < \beta_h \le b)\pi(\beta|Y)\, d\beta \\ &\doteq \sum_{j=1}^{t} I(a < \beta_h^j \le b)\tilde{q}_j\end{aligned}$$

where $I(A) = 1$ if A occurs, and 0 otherwise.

Table 2 presents posterior means, standard deviations, and the 5th and 95th percentiles for the regression coefficients from the trauma data using our prior and for the diffuse prior $\pi(\beta) = 1$. The posterior means are Bayes estimates of the regression coefficients. The maximum likelihood estimates and asymptotic standard errors are also given for comparison. Using our prior, $\Pr(\beta_1 > 0|Y) > .99$, which suggests that the coefficient of ISS is positive. As naively expected, the RTS coefficient is negative. AGE, TI, and their interaction show little effect when fitted last. Similar conclusions were found using the maximum likelihood estimates and under the diffuse prior, albeit the estimated intercept using our prior is substantially larger.

The central 90% posterior intervals for individual regression coefficients based on our prior are approximately 3/4 as wide as the intervals based on a diffuse prior, so any inferences about the regression coefficients will not be overly sensitive to the prior information. Keep in mind, however, that prediction, and not estimation, is the primary goal in a survival model. A more careful analysis of the effect of our prior on predictive inferences is given in Section 2.7.

2.5 Case Deletion Diagnostics

Johnson (1985) introduced diagnostics based on case deletions for estimating probabilities in logistic regression. Our purpose is to detect cases that noticeably affect our inferences. For example, we would want to know if the predictive probability of death changes radically upon deleting one case. It is also relevant to examine the effect of the prior on posterior inferences. Our prior is of the data augmentation type, so it is possible to look at the effect of deleting a "prior observation"; see Section 2.7.

The effect of case deletion on the posterior of β is easily formulated. Let $Y_{(i)}$ denote the data without y_i. The likelihood for β based on all the data except y_i is $L(\beta|Y_{(i)}) = L(\beta|Y)/L(\beta|y_i)$, so

$$\pi(\beta|Y_{(i)}) = \frac{L(\beta|Y_{(i)})\pi(\beta)}{\int L(\beta|Y_{(i)})\pi(\beta)\, d\beta} = \frac{\pi(\beta|Y)/L(\beta|y_i)}{\int \pi(\beta|Y)/L(\beta|y_i)\, d\beta}. \tag{2.4}$$

Table 2. Fitted Trauma Model

		Posterior summaries				Maximum likelihood	
Prior	*Variable*	*Estimate (Mean)*	*Std. error (Std. dev.)*	*.05%*	*.95%*	*Estimate*	*Std. error*
BCJ	Intercept	−1.79	1.10	−3.54	.02	−2.73	1.62
	ISS	.07	.02	.03	.10	.08	.03
	RTS	−.60	.14	−.82	−.37	−.55	.17
	AGE	.05	.01	.03	.07	.05	.01
	TI	1.10	1.06	−.66	2.87	1.34	1.33
	AGE × TI	−.02	.03	−.06	.03	−.01	.03
Diffuse	Intercept	−2.81	1.60	−5.34	−.18		
	ISS	.09	.03	.04	.13		
	RTS	−.59	.17	−.86	−.32		
	AGE	.06	.02	.03	.09		
	TI	1.46	1.36	−.79	3.69		
	AGE × TI	−.01	.03	−.07	.05		

If we renormalize the weights

$$\tilde{q}_{j(i)} = \frac{\tilde{q}_j / L(\beta^j | y_i)}{\sum_{k=1}^{t} \tilde{q}_k / L(\beta^k | y_i)}$$

then the distribution taking values β^j with probability $\tilde{q}_{j(i)}$ gives a discrete approximation to the posterior (2.4). Expectations with respect to $\pi(\beta | Y_{(i)})$ are evaluated using this approximate distribution.

The symmetric Kullback–Leibler (KL) divergence is used to measure the discrepancy between full and reduced data predictive distributions. The predictive distribution for a single trial is Bernoulli. The symmetric KL divergence between two Bernoulli distributions with probabilities p and q is $J(p, q) \equiv (p - q) \log(p(1 - q)/(1 - p)q)$.

As in Johnson (1985) we define a symmetric predictive divergence diagnostic for predicting new observations at the original data locations when case i is deleted as

$$D_i^p \equiv \sum_{j=1}^{n} J(p(y = 1 | Y, x_j), p(y = 1 | Y_{(i)}, x_j)). \quad (2.5)$$

Here, $p(y = 1 | Y, x)$ is the predictive probability of success from all of the data as defined in (2.3), and $p(y = 1 | Y_{(i)}, x)$ is the predictive probability of success based on all of the data except case i:

$$p(y = 1 | Y_{(i)}, x) = \int r(x'\beta) \pi(\beta | Y_{(i)}) \, d\beta \doteq \sum_{j=1}^{t} r(x'\beta^j) \tilde{q}_{j(i)}.$$

Alternatively, a symmetric predictive divergence diagnostic D_i^f for predicting observations at an arbitrary set of locations, say x_j^f, $j = 1, \ldots, r$ can be defined by replacing the x_js in (2.5) with predictor covariate vectors x_j^f. A large value of D_i^p or D_i^f indicates that deletion of case i results in different predictive probabilities than if it were retained, possibly resulting in different inferences or decisions.

Computing the D_i^ps we found cases 52 and 232 to be most influential. Case 52 is a 66 year old man who had little wrong with him (ISS = 9, RTS = 7.84) with a penetration injury who died. Case 232 is a very sick and damaged (RTS = 2.19, ISS = 50) 50 year old person with a blunt injury who managed to survive. An ISS score of 50 is characteristic of a person who has very severe injuries to two different parts of the body. The actual statistics are $D_{52}^p = .46$ and $D_{232}^p = .41$, with the next highest value being $D_{173}^p = .25$. Figure 4 contains an index plot of the difference in the predictive probabilities of death, $p(y = 1 | Y, x_j) - p(y = 1 | Y_{(52)}, x_j)$. These probabilities depend on the specified prior. Note that, having deleted a case in which a relatively healthy person died, most of the probability differences are very near 0, but slightly positive, for example, most people's probabilities of death have decreased. Moreover, all of the changes in probabilities are relatively small. Deletion of case 232 seems to change the regression coefficients even less, and would seem to have even less effect on the fitted probabilities.

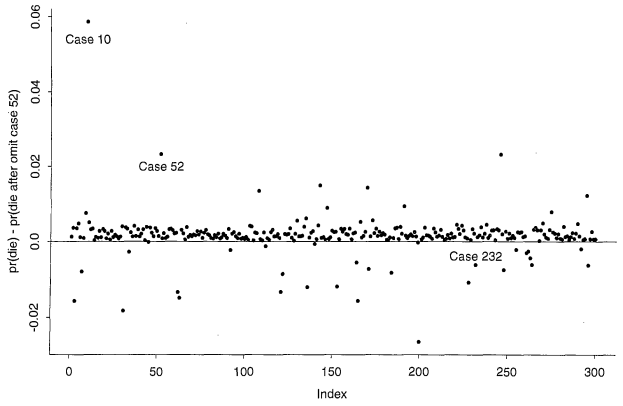

Figure 4. Change in Predictive Probability of Dying Upon Omitting Case 52.

One can also examine case deletion diagnostics for the estimation of β. As in Johnson and Geisser (1985) one can define the KL symmetric divergence between the full posterior of β and the posterior of β after deleting case i. However, theoretical and empirical observations (Johnson 1985) lead us to expect similarity between these measures and the predictive divergences.

2.6 Link Selection

We now allow the Bayesian paradigm to indicate which of the three link function models is most appropriate for the data: logistic (M_1), probit (M_2), or complementary log–log (M_3). Bayes factors for comparing models M_j and M_k are numbers BF_{jk} such that

$$\frac{P(M_j | Y)}{P(M_k | Y)} = [BF_{jk}] \frac{P(M_j)}{P(M_k)}.$$

The Bayes factor is the multiplier that changes the prior odds for the models into the posterior odds. It is a simple application of Bayes's theorem to show that

$$BF_{jk} = \frac{p(Y | M_j)}{p(Y | M_k)}$$

where $p(Y | M_i)$ is the marginal probability of obtaining Y from model M_i. Computing $p(Y | M)$ for a model M involves integrating the corresponding likelihood function with respect to the induced prior on β for that model.

For the trauma data the Bayes factors under our prior are $BF_{21} = 1.05$, $BF_{13} = 20.72$, and thus $BF_{23} = BF_{21}/BF_{31} = BF_{21}BF_{13} = 1.05(20.72) = 21.83$. This suggests that the complementary log–log model is inappropriate, but there is little to choose from between the logistic and probit models. If the prior odds for the probit versus logit models are 1, the posterior odds are merely 1.05.

The model selection computations were performed by sampling from the prior distribution. We sample the vectors $\tilde{p}$ and solve $\tilde{p} = R(\tilde{X}\beta)$ to obtain samples of β. Sampling $\tilde{p}$ is easy with our prior because the $\tilde{p}_i$s have independent Beta distributions. Given a sample $\beta_{\#}^j, j = 1, \ldots, v$ from the prior, for model M

$$p(Y | M) \doteq \frac{1}{v} \sum_{j=1}^{v} L(\beta_{\#}^j | Y).$$

We used $v = 10,000$ in our calculations.

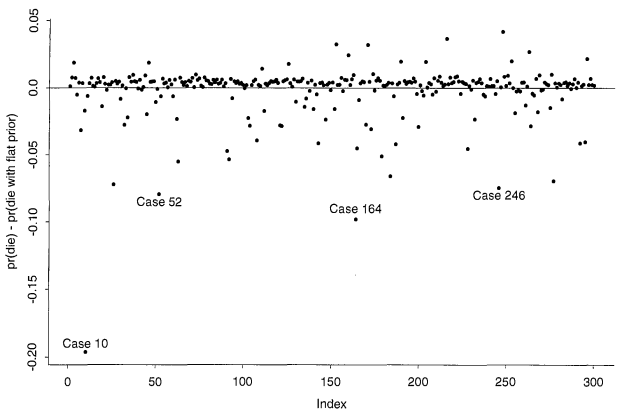

Figure 5. Change in Predictive Probability Between Proper and Flat Priors.

One can also form a model selection case deletion diagnostic by computing $BF_{jk(i)} = p(Y_{(i)}|M_j)/p(Y_{(i)}|M_k)$, where $p(Y_{(i)}|M) = \int L(\beta|Y_{(i)}, M)\pi(\beta|M)\, d\beta$.

2.7 Sensitivity of Predictions to the Prior

The sensitivity of posterior inferences to the choice of the prior can be evaluated by recalculating posterior summaries based on alternative priors. To examine sensitivity to the prior specifications, we considered case deletions of the "prior observations." We repeated the methods of Section 2.5 except that, rather than deleting observation i, we left in all of the data, but changed the prior by deleting one of the six prior observations. The six values of the symmetric predictive divergences were, in order, .00, .20, .12, .01, .01, .03. These can be compared in magnitude to the D_i^ps. The largest D_i^ps were around .45; the values from the prior are all much smaller. We also constructed plots similar to Figure 4 for each of these six deletions. The plots showed substantially smaller changes than Figure 4, with prior observations 2 and 3 causing the largest changes.

We also compared the predictions from our prior to the predictions based on the diffuse prior $\pi(\beta) = 1$. Figure 5 gives an index plot of the difference in the predictive probabilities of death using the two priors. The symmetric predictive divergence was 1.14. The cumulative effect of eliminating the prior information at each of the six covariate conditions is large relative to the effect of eliminating any individual prior observation. Although the mean absolute difference between predictive probabilities is only .01, several cases, notably 10, 52, 164, and 246, have much lower predictive probabilities of death under our prior. A closer look at the data shows that many of these cases have relatively high leverage. In essence, our prior has a large effect on the predictive probabilities for cases with extreme covariate conditions, but little effect on cases near the center of design space. This result is not surprising because our prior locations are fairly extreme within the design space.

3. POSTERIOR COMPUTATIONS AND SAMPLE SIZE CALCULATION

Gelman, Carlin, Stern, and Rubin (1995, chap. 9–11) provide a nice general discussion of simulation methods and numerical methods for approximating posterior inferences. We have focused on using discrete approximations to the posterior distribution. Specifically, we used a variant of importance sampling. For this one should choose an importance density function $g(\beta)$ that is similar in shape to the known kernel of the posterior $L(\beta|Y)\pi(\beta)$. Importance sampling can be inefficient when the shape of the posterior density is poorly matched by $g(\beta)$. From the distribution with density $g(\beta)$ sample $\beta^1, \ldots, \beta^t$. For $i = 1, \ldots, t$ compute the weights $q_i = q(\beta^i) = L(\beta^i|Y)\pi(\beta^i)/g(\beta^i)$ and $\tilde{q}_i = q_i / \sum_{j=1}^t q_j$. The discrete approximation to the posterior distribution takes values β^i with probability $\tilde{q}_i$. Geweke (1989) discusses the efficiency and accuracy of this method, and gives diagnostic procedures. The tails of the importance function should decay less rapidly than those of the posterior to avoid having individual samples in the tail of the importance distribution with extremely large weights.

A natural choice for the importance density $g(\beta)$ is a multivariate Student t density with s degrees of freedom with location equal to the posterior mode β_M and dispersion proportional to $\Sigma(\beta_M)$, the asymptotic posterior covariance matrix evaluated at the mode. For DAPs like (2.2), these posterior quantities can be obtained from standard GLM software because they are just the MLE and asymptotic covariance matrix as computed from the augmented data. For non-DAP priors β_M and $\Sigma(\beta_M)$ must be computed using specialized software. Unfortunately, no single method of picking $g(\beta)$ can be expected to work well for every prior and link function. In our example we used a t with 6 degrees of freedom to ensure that the tails were heavy. We allowed for skewness in the posterior distribution by "splitting" the t distribution at the posterior mode. The optimal splitting parameters were obtained using Geweke's (1989) algorithm.

We decided on the size of the importance sample by first generating a pilot sample of 2,500 observations. Because prediction was a primary goal we decided that the average coefficient of variation in the estimated survival probabilities should not exceed 5%. Our pilot study indicated that a total sample size of approximately 8,000 would suffice. We sampled 10,000 observations, which gave an estimated average coefficient of variation for the parameters of interest of 4.1%. The estimated median coefficient of variation was 1.2%.

We should mention that our importance sampling approach can be inaccurate for approximating case deletion diagnostics if some of the renormalized weights $\tilde{q}_{j(i)}$ used in the approximate posterior $\pi(\beta|Y_{(i)})$ are too large. In such cases the case-deleted inferences are essentially based on only a few posterior samples having large weights. An importance resampling approach, as described in Gelman et al. (1995, chap. 9) would avoid this problem, but this was not needed in our analysis.

4. CONCLUDING REMARKS

We have presented a simple approach to a fully Bayesian analysis of binomial regression. Primary features include specification of a prior that can be elicited easily, model selection, diagnostic checking, and numerical methods that

are simple and accurate. In the interest of brevity we have not discussed global model checking, but we have found the procedures of Box (1980) to be useful. Other generalized linear model problems can be handled similarly; see BCJ. Statisticians no longer need to rely on asymptotics or avoid Bayesian methods in these settings because of difficulties with specifying a prior, model validation, or intractable integrations.

The recent texts by Carlin and Louis (1996), Gelman et al. (1995), and Gilks, Richardson, and Spiegelhalter (1996) examine a variety of complex modeling problems that can be handled easily using Bayesian methods. More theoretical aspects of the Bayesian approach can be found in Berger (1985) and Bernado and Smith (1994). A standard reference for Bayesian prediction is Aitchison and Dunsmore (1975).

[Received June 1993. Revised March 1997.]

REFERENCES

Aitchison, J., and Dunsmore, I. R. (1975), *Statistical Prediction Analysis*, Cambridge: Cambridge University Press.

Baker, S. P. et al. (1974), "The Injury Severity Score: A Method for Describing Patients with Multiple Injuries and Evaluating Emergency Care," *Journal of Trauma*, 14, 187–196.

Bedrick, E. J., Christensen, R., and Johnson, W. (1996), "A New Perspective on Priors for Generalized Linear Models," *Journal of the American Statistical Association*, 91, 1450–1460.

Berger, J. O. (1985), *Statistical Decision Theory and Bayesian Analysis* (2nd ed.), New York: Springer-Verlag.

Bernado, J. M., and Smith, A. F. M. (1994), *Bayesian Theory*, New York: Wiley.

Box, G. E. P. (1980), "Sampling and Bayes' Inference in Scientific Modelling and Robustness," *Journal of the Royal Statistical Society*, Ser. A, 143, 383–404.

Carlin, B. P., and Louis, T. A. (1996), *Bayes and Empirical Bayes Methods for Data Analysis*, New York: Chapman and Hall.

Champion, H. R., Sacco, H. J., Copes, W. S. et al. (1989), "A Revision of the Trauma Score," *Journal of Trauma*, 29, 623–629.

Champion, H. R., Sacco, H. J., and Copes, W. S. (1991), "Trauma Scoring in Trauma," in *Trauma*, ed. E. E. Moore, K. L. Mattox, and D. V. Feliciano, Norwalk, CT: Appleton and Lange, pp. 47–65.

Dellaportas, P., and Smith, A. F. M. (1993), "Bayesian Inference for Generalized Linear Models and Proportional Hazards via Gibbs Sampling," *Applied Statistics*, 42, 443–459.

Geisser, S. (1982), "Aspects of the Predictive and Estimative Approaches to the Determination of Probabilities," *Biometrics Supplement: Current Topics in Biostatistics and Epidemiology*, 38, 75–93.

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

Gilks, W. R., Richardson, S., and Spiegelhalter, D. (eds.) (1996), *Practical Markov Chain Monte Carlo*, New York: Chapman and Hall.

Grieve, A. P. (1988), "A Bayesian Approach to the Analysis of LD50 Experiments," in *Bayesian Statistics 3*, eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, Oxford: Oxford University Press, pp. 617–630.

Geweke, J. (1989), "Bayesian Inference in Econometric Models Using Monte Carlo Integration," *Econometrica*, 57, 1317–1339.

Johnson, W. (1985), "Influence Measures for Logistic Regression: Another Point of View," *Biometrika*, 72, 59–65.

Johnson, W., and Geisser, S. (1985), "Estimative Influence Measures for the Multivariate General Linear Model," *Journal of the Statistical Planning and Inference*, 11, 33–56.

Lee, P. M. (1989), *Bayesian Statistics: An Introduction*, London: Edward Arnold.

Leonard, T. (1972), "Bayesian Methods for Binomial Data," *Biometrika*, 59, 581–589.

Osler, T. (1993), "Injury Severity Scoring: Perspectives in Development and Future Directions," *American Journal of Surgery*, 165 (2A, suppl.), 43S–51S.

Smith, A. F. M., Skene, A. M., Shaw, J. E. H., Naylor, J. C., and Dransfield, M. (1985), "The Implementation of the Bayesian Paradigm," *Communications in Statistics—Theory and Methods*, 14, 1079–1102.

Tierney, L., and Kadane, J. (1986), "Accurate Approximations for Posterior Moments and Marginal Densities," *Journal of the American Statistical Association*, 81, 82–86.

Tsutakawa, R. K. (1975), "Bayesian Inference for Bioassay," Technical Report 52, University of Missouri, Columbia, Dept. of Statistics.

Tsutakawa, R. K., and Lin, H. Y. (1986), "Bayesian Estimation of Item Response Curves," *Psychometrika*, 51, 251–267.

Zellner, A., and Rossi, P. E. (1984), "Bayesian Analysis of Dichotomous Quantal Response Models," *Journal of Econometrics*, 25, 365–393.