A Bayesian Analysis of a Poisson Random Effects Model for Home Run Hitters

JAMES ALBERT*

The problem of interest is to estimate the home run ability of 12 great major league players. The usual career home run statistics are the total number of home runs hit and the overall rate at which the players hit them. The observed rate provides a point estimate for a player's "true" rate of hitting a home run. However, this point estimate is incomplete in that it ignores sampling errors, it includes seasons where the player has unusually good or poor performances, and it ignores the general pattern of performance of a player over his career. The observed rate statistic also does not distinguish between the peak and career performance of a given player. Given the random effects model of West (1985), one can detect aberrant seasons and estimate parameters of interest by the inspection of various posterior distributions. Posterior moments of interest are easily computed by the application of the Gibbs sampling algorithm (Gelfand and Smith 1990). A player's career performance is modeled using a log-linear model, and peak and career home run measures for the 12 players are estimated.

KEY WORDS: Gibbs sampling; Log-linear model; Outliers; Overdispersion; Simulation.

1. INTRODUCTION

One of the most dramatic aspects of a baseball game is the home run. Throughout the last century, the American public has had a fascination with the home run and with the hitters who possess the talent to hit a high frequency of home runs. Indeed, many historians have written biographies of the great home run hitters in baseball, such as Babe Ruth, Henry Aaron, and Mickey Mantle, and attempted to rank the greatest sluggers of all time. [See James (1988) and Thorn and Palmer (1984) for rankings of baseball hitters with respect to various criteria.]

In rating home run hitters, two measures are typically used. The most naive measure is the number of home runs hit during the player's major league career. Using data from Neft and Cohen (1985), Table 1 lists 12 of the great home run hitters using this criterion, with Henry Aaron and Babe Ruth leading the list. Henry Aaron is generally perceived by the American public as the greatest home run hitter since he hit the greatest number in a career. Although this statistic measures a player's consistency in hitting a high frequency of home runs over a long career, it may not measure a player's ability to hit a home run at a given plate appearance.

*James Albert is Professor, Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, OH 43403.

An alternative statistic for measuring a player's home run ability divides the total number of home runs in a career by the total number of official at-bats. This statistic is the (observed) rate, which can be viewed as an estimate of the "true" rate (over a career) at which a player hit home runs. Table 1 also ranks these 12 hitters with respect to this criterion. Note that, although Aaron hit more home runs than Ruth, Ruth hit them at a higher rate. In baseball statistics books such as Siwoff, Hirdt, Hirdt, and Hirdt (1991), this home run rate statistic is usually expressed as the number of at-bats for each home run hit and these rates are also listed in Table 1. In this article, the rate home runs/at-bats will be used since it corresponds to the usual rate parameter in the sampling models that will be considered.

Both of these home run criteria can be used to estimate a hitter's talent or ability to hit a home run during a plate appearance. Clearly, the rate statistic HR/AB is better than the total number of home runs HR in measuring this true rate. However, this rate statistic is only an estimate of a player's true rate, and differences between two players' observed rates may entirely reflect sampling error. Using a Poisson sampling model and a noninformative prior placed on a player's true rate, Section 2.1 derives the rate estimate and an associated standard error. These standard errors can be used to assess which players' true rates are significantly different.

A second criticism of this rate statistic, which is the focus of this article, is that no adjustment or allowance is made for particular seasons in which a player's home run performance is unusually good or poor. A player can be "hot" and hit a high number of home runs for one particular season. Good examples of this are Roger Maris's 61-home run season in 1961 and, more recently, Cecil Fielder's 51-home run season in 1990. Maris never hit close to 61 home runs in any other season in his

Table 1. Ranking of 12 Hitters With Respect to Two Home Run Criteria: Career Home Runs and Home Run Rate

Hitter	Career home runs (HR)	Career at-bats (AB)	Hitter	Home run rate (HR/AB)	(AB/HR)
H. Aaron	755	12,364	B. Ruth	.0851	11.75
B. Ruth	714	8,389	R. Kiner	.0709	14.11
W. Mays	660	10,881	H. Killebrew	.0703	14.22
H. Killebrew	573	8,147	T. Williams	.0676	14.80
M. Schmidt	548	8,352	M. Schmidt	.0662	15.12
M. Mantle	536	8,102	M. Mantle	.0657	15.23
J. Foxx	534	8,134	J. Foxx	.0656	15.24
W. McCovey	521	8,173	H. Greenberg	.0638	15.69
T. Williams	521	7,696	W. McCovey	.0637	15.69
L. Gehrig	493	8,001	L. Gehrig	.0616	16.22
R. Kiner	369	5,205	H. Aaron	.0611	16.38
H. Greenberg	331	5,192	W. Mays	.0607	16.49

career; in particular, he hit only 39 and 33 in the neighboring seasons 1960 and 1962. In addition, due to injuries or other reasons, a player's home run performance can be unusually poor during certain seasons. Mickey Mantle is a good example of a great player whose injuries had a detrimental effect on his hitting performance.

The rate statistic can be very sensitive to unusually good or poor home run performances during particular seasons. To correct this problem, Poisson random effects models are introduced in Section 2.2 to accommodate the aberrant seasons. Suppose that the player's career consists of p seasons. Then one random effects model, initially used by West (1985), adds p additional scale parameters to the Poisson model. Using a vague prior on the rate parameter and the individual scale parameters, the objective is to obtain marginal posterior distributions of all $p + 1$ parameters. The posterior densities of the scale parameters are of interest since they will aid in identifying seasons that are inconsistent with the main body of data. The posterior density of the rate parameter accommodates the unusual observations by downweighting these particular seasons.

The posterior calculations using this new model appear formidable in this problem due to the high dimensionality of the posterior. West (1985) only obtained modal estimates for the scale parameters. However, using the Gibbs sampler algorithm, all of these one-dimensional marginal posterior distributions can be found by simulation. Section 3 outlines this simulation algorithm and shows that this method is easily implemented for this random effects model.

The simulation algorithm is illustrated for Henry Aaron's home run data in Section 4. Using this algorithm on 12 leading players, a new rating of the players is made that accommodates unusually good or poor seasons.

This random effects model improves over the Poisson model, but it can be made more realistic. The above model assumes that a player's true home run rate remains constant over the life of his career. This is unsatisfactory in that it is well-known by baseball experts that the typical player's batting ability increases as he matures, peaks about the middle of his career, and slowly decreases until retirement. Therefore, a better model for a player's true home run rate is the quadratic log-linear model

$$\log \text{rate}_i = \beta_0 + \beta_1 i + \beta_2 i^2,$$

where rate_i represents the true rate during the ith season, $i = 1, \ldots, p$. Morris (1983) modeled Ty Cobb's seasonal batting averages using a similar quadratic model. Section 5 outlines the application of the Gibbs sampler for this more general model. For this model, one can distinguish between the peak home run performance and career performance of a player. Both performances can be measured by particular functions of the fitted rates $\{\exp(\beta_0 + \beta_1 i + \beta_2 i^2)\}$ and this section discusses how to obtain posterior moments for these functions of interest.

In Section 6, we summarize this new quadratic random effects model by rating the 12 premier home run hitters with respect to career and peak performance. Since this article is written using a Bayesian perspective, this section relates this work with non-Bayesian approaches for identifying outliers.

2. MODELING HOME RUN DATA

2.1 The Poisson Sampling Model

For a particular player, let y_i denote the proportion of home runs hit in t_i at-bats during the ith season. A simple model is that $t_i y_i$, the number of home runs hit, is binomial (t_i, p), where p is the probability of hitting a home run during a given plate appearance. Since p is small, it is reasonable to approximate the distribution of $t_i y_i$ by a Poisson distribution with mean $t_i\lambda$, where λ is the player's true home run rate over his career. The sampling density of y_i is given by

$$f(y_i|\lambda) = \frac{e^{-t_i\lambda}(t_i\lambda)^{t_i y_i}}{(t_i y_i)!}, \qquad t_i y_i = 0, 1, 2, \ldots. \quad (1)$$

If the observed proportions $y_1, \ldots, y_p$ are assumed independent, then the likelihood of λ is given by

$$L(\lambda) = \prod_{i=1}^{p} f(y_i|\lambda) = e^{-\lambda\Sigma t_i}\lambda^{\Sigma t_i y_i}\prod_{i=1}^{p}\frac{t_i^{t_i y_i}}{(t_i y_i)!}. \quad (2)$$

If λ is assigned the noninformative prior $\pi(\lambda) = \lambda^{-1}$, then the posterior density of λ is given by

$$\pi(\lambda|\mathbf{y}) = Ce^{-\lambda\Sigma t_i}\lambda^{\Sigma t_i y_i - 1}, \qquad \lambda > 0, \quad (3)$$

where $\mathbf{y} = (y_1, \ldots, y_p)$ and C is a generic proportionality constant.

The posterior density (3) is of the gamma form with shape parameter $\alpha = \Sigma\, t_i y_i$ and scale parameter $\varphi = \Sigma\, t_i$. The posterior mean and standard deviation of λ are given respectively by $E[\lambda|\mathbf{y}] = \Sigma\, t_i y_i/\Sigma\, t_i$ and $\text{SD}[\lambda|\mathbf{y}] = \sqrt{\Sigma\, t_i y_i}/\Sigma\, t_i$. Note that the posterior mean is the rate statistic given in Table 1; the posterior standard deviation can be interpreted as a standard error of this rate statistic.

2.2 Poisson Random Effects Models

The Poisson model is not a good fit to the home run data for the 12 sluggers. This model assumes that $E[y_i] = \text{var}[\sqrt{t_i}y_i] = \lambda$. If one plots the sample mean of $\{y_i\}$ against the sample variance of $\{\sqrt{t_i}y_i\}$ for each of the 12 hitters, one sees that the variance estimate is consistently higher than the mean estimate. Thus, the variability of the home run data is generally much higher than predicted by the Poisson model. Using the terminology of Nelder and McCullagh (1989), the data are overdispersed and it is of interest to fit an alternative model that can accommodate this extra variability.

One general method of modeling extra variability uses mixtures. As above, suppose that $y_i t_i$ is distributed Poisson $(t_i\lambda)$, but λ is random and assigned the gamma

distribution

$$g(\lambda) = \frac{\lambda^{\eta_i m - 1} e^{-\lambda \eta_i} \eta_i^{\eta_i m}}{\Gamma(\eta_i m)}. \quad (4)$$

One can show that the marginal distribution of y_i is given by the negative binomial form

$$f^*(y_i|m, \eta_i) = \frac{\eta_i^{\eta_i m} t_i^{y_i t_i} \Gamma(\eta_i m + y_i t_i)}{(y_i t_i)!(\eta_i + t_i)^{\eta_i m + y_i t_i} \Gamma(\eta_i m)}. \quad (5)$$

The mean and variance of y_i are given by

$$E[y_i] = m, \quad \text{var}[y_i] = \frac{m}{t_i}\left(\frac{\eta_i + t_i}{t_i}\right). \quad (6)$$

In this setting, m would represent a batter's true rate and η_i is an additional season-specific parameter that can model the extra variation in the data.

The above negative binomial model appears suitable for accommodating great fluctuations in a batter's seasonal home run data. Due to the gamma functions in the likelihood (5), however, it cannot be fit using standard statistics software routines. In this article, an alternative model is used, based on the addition of a scale parameter to the Poisson density (1). This alternative random effects model is attractive in that it is very easy to fit and can be easily generalized to model an arbitrary log-linear model for the rate parameter.

To motivate the construction of this alternative model, first note that if we approximate $(t_i y_i)!$ in (1) using Stirling's formula

$$(t_i y_i)! \approx \sqrt{2\pi}\,(t_i y_i)^{t_i y_i + 1/2} e^{-t_i y_i},$$

we obtain the approximate Poisson sampling density

$$f(y_i|\lambda) \approx \frac{1}{\sqrt{2\pi}\sqrt{t_i y_i}} \exp\{-D(y_i, \lambda)/2\}, \quad (7)$$

where $D(y_i, \lambda)$ is the deviance

$$\begin{aligned} D(y_i, \lambda) &= -2[\log f(y_i|\lambda) - \log f(y_i|y_i)] \\ &= -2\{t_i[y_i - \lambda - y_i \log(y_i/\lambda)]\}. \end{aligned}$$

This approximate density is referred as a quasi-likelihood (QL) in Nelder and McCullagh (1989) since it does not correspond to a particular sampling density.

Note that (7) generalizes the normal density with $(y_i - \lambda)^2$ in the exponent replaced by the Poisson deviance $D(y_i, \lambda)$. Then, analogous with the normal density, one can then generalize the QL sampling model by adding a scale parameter γ_i:

$$g(y_i|\lambda, \gamma_i) = \frac{\gamma_i^{1/2}}{\sqrt{2\pi}\sqrt{t_i y_i}} \exp\{-\gamma_i D(y_i, \lambda)/2\}. \quad (8)$$

For the Poisson sampling model, $E[y_i|\lambda] = \lambda$ and $\text{var}[y_i|\lambda] = \lambda/t_i$. The QL sampling model (8) extends the Poisson model by the more general variance structure $\text{var}[y_i|\lambda, \gamma_i] = \lambda/t_i\gamma_i$. Note that this model gives the same variance structure as the Poisson-gamma mixture density (5). Both the negative binomial density and the Poisson QL density are common models for fitting overdispersed count data. [See Lawless (1987), Thall (1988), and Albert and Pepple (1989), for illustrations using these models.)]

If $y_1, \ldots, y_p$ are assumed independent and given QL models with respective scale parameters $\gamma_1, \ldots, \gamma_p$, then the joint likelihood of $(\lambda, \{\gamma_i\})$ is given by

$$L(\lambda, \gamma_1, \ldots, \gamma_p) = \exp\left\{-\frac{1}{2}\sum_{i=1}^{p} \gamma_i D(y_i, \lambda)\right\} \prod_{i=1}^{p} \gamma_i^{1/2}. \quad (9)$$

To complete the model specification, we assume the $p + 1$ parameters are independent—we assign the Poisson rate parameter λ the usual noninformative prior λ^{-1} and assume $\gamma_1, \ldots, \gamma_p$ are iid from the gamma density given by

$$\pi(\gamma_j) = \gamma_j^{\nu/2 - 1} e^{-\gamma_j \nu/2} (\nu/2)^{\nu/2}/\Gamma(\nu/2). \quad (10)$$

The joint posterior density of all $p + 1$ parameters is given by

$$\begin{aligned} \pi(\lambda, \gamma_1, \ldots, \gamma_p|y) &= C \exp\left\{-\frac{1}{2}\sum_{i=1}^{p} \gamma_i[D(y_i, \lambda) + \nu]\right\} \\ &\quad \times \lambda^{-1} \prod_{i=1}^{p} \gamma_i^{(\nu+1)/2 - 1}. \quad (11) \end{aligned}$$

An equivalent representation of (11) is

$$\begin{aligned} &C\lambda^{\sum_{i=1}^{p} \gamma_i t_i y_i - 1} \exp\left\{-\lambda \sum_{i=1}^{p} \gamma_i t_i\right\} \\ &\quad \times \exp\left\{\sum_{i=1}^{p} \gamma_i (t_i[-y_i \log y_i + y_i] - \nu/2)\right\}. \quad (12) \end{aligned}$$

Note that the posterior density (11) is singular since there are $p + 1$ parameters with p data points. This will be not be a problem, since the interest is in the marginal posterior densities of the parameters, which are proper distributions. In addition, note that if one was solely interested in the marginal posterior density of λ, then one could integrate out the scale parameters $\gamma_1, \ldots, \gamma_p$, obtaining

$$\pi(\lambda|\mathbf{y}) = C\lambda^{-1} \prod_{i=1}^{p} (\nu + D(y_i, \lambda))^{(\nu+1)/2 - 1}. \quad (13)$$

The marginal density (13) is the posterior density of the rate parameter λ when $y_1, \ldots, y_p$ are a random sample from the sampling density $(\nu + D(y_i, \lambda))^{(\nu+1)/2 - 1}$. This sampling density can be viewed as generalization of a t density with the normal deviance $(y_i - \lambda)^2$ in the t density form replaced by the Poisson deviance $D(y_i, \lambda)$. In this problem, the representation (11) will be used, since all $p + 1$ parameters are of interest and it is convenient to simulate from the distribution when expressed in this multivariate form.

In practice, one needs to choose a value of the gamma shape parameter ν in (10). In examples, $\nu = 10$ is used, reflecting a vague prior belief about the location of the scale parameters $\{\gamma_i\}$. In Section 4, the sensitivity of the posterior analysis with respect to this parameter is briefly

investigated. One could regard the parameter ν as unknown, assign it a prior, and estimate it using the resulting posterior distribution computed via Gibbs sampling. The parameter ν can be viewed as a measure of the consistency of a player over his career. We will not pursue this estimation problem here, since the main emphasis is on the estimation of a player's home run rate.

3. SIMULATION OF ALL MARGINAL POSTERIOR DISTRIBUTIONS VIA THE GIBBS SAMPLER

Gelfand and Smith (1990) and Gelfand, Hills, Racine-Poon, and Smith (1990) have recently illustrated the use of a new method for simulating a multivariate posterior distribution. This sampling algorithm, the Gibbs sampler, has been shown to be successful in image restoration (Geman and Geman 1984), but its statistical applications have not been recognized until recently. For two random variables X and Y, let $[X, Y]$, $[X|Y]$ and $[X]$ denote joint, conditional, and marginal posterior distributions. In this setting, it is of interest to simulate from the $p + 1$ dimensional posterior distribution $[\lambda, \gamma_1, \ldots, \gamma_p]$. Although this distribution (11) is complicated, the fully conditional distributions (the distributions of a single parameter conditional on all the remaining parameters) are tractable. In particular, from (12), note that the posterior distribution of λ conditional on $\{\gamma_i\}$ has the gamma form

$[\lambda|\gamma_i, \ldots, \gamma_p]$ distributed

$$\text{gamma}\left(\sum_{i=1}^{p} \gamma_i t_i y_i, \sum_{i=1}^{p} \gamma_i t_i\right). \quad (14)$$

In addition, using (11), the posterior distributions of $\gamma_1, \ldots, \gamma_p$, conditional on λ, are independent with

$[\gamma_i|\lambda]$ distributed

$$\text{gamma}\left(\frac{\nu + 1}{2}, \frac{D(y_i, \lambda) + \nu}{2}\right). \quad (15)$$

The Gibbs sampler simulates the parameters using the entire set of fully conditional distributions. To implement this simulation method, one starts with initial values of the $p + 1$ parameters $(\lambda^{(0)}, \gamma_1^{(0)}, \ldots, \gamma_p^{(0)})$. In this problem, convenient starting values are the Poisson model estimates $\gamma_i^{(0)} = 1$, $i = 1, \ldots, p$ and $\lambda^{(0)} = \Sigma\, y_i t_i / \Sigma\, t_i$. Then letting $\boldsymbol{\gamma}^{(t)} = (\gamma_1^{(t)}, \ldots, \gamma_p^{(t)})$, one simulates in turn $\boldsymbol{\gamma}^{(1)}$ from $[\boldsymbol{\gamma}|\lambda^{(0)}]$, $\lambda^{(1)}$ from $[\lambda|\boldsymbol{\gamma}^{(1)}]$, $\boldsymbol{\gamma}^{(2)}$ from $[\boldsymbol{\gamma}|\lambda^{(1)}]$, $\lambda^{(2)}$ from $[\lambda|\boldsymbol{\gamma}^{(2)}]$, and so on. After t cycles of the sequence

$$\{\boldsymbol{\gamma}^{(k)} \sim [\boldsymbol{\gamma}|\lambda^{(k-1)}],\ \lambda^{(k)} \sim [\lambda|\boldsymbol{\gamma}^{(k)}]\}, \quad (16)$$

we obtain one simulated value $(\lambda^{(t)}, \gamma^{(t)})$ from the joint distribution $[\lambda, \boldsymbol{\gamma}]$.

One convenient and efficient method of performing the simulation is to start with the above initial values and run the sampler for a large number of cycles m. (This method will be called "one long chain.") If one plots the $(\lambda^{(t)}, \boldsymbol{\gamma}^{(t)})$ against t, then the values appear to settle down and converge quickly to the posterior $[\lambda, \boldsymbol{\gamma}]$. Thus we take as our simulated sample the entire set $\{(\lambda^{(k)}, \boldsymbol{\gamma}^{(k)}),\ k = 1, \ldots, m\}$. In our examples, a few hundred cycles appeared sufficient to understand the behavior of these random effects models. These simulated values are used in the calculation of posterior expectations $E[g(\lambda, \gamma_1, \ldots, \gamma_p)|y]$. For example, the mean $E[\lambda|y]$ is estimated by the sample mean of the simulated values $\Sigma_{k=1}^{m} \lambda^{(k)}/m$. Note that this sample will be correlated, since values of $(\lambda^{(k)}, \boldsymbol{\gamma}^{(k)})$ and $(\lambda^{(k+l)}, \boldsymbol{\gamma}^{(k+l)})$ will be positively correlated for small values of the lag l. The only difficulty with this correlation is that is harder to compute a numerical standard error for the estimated posterior expectations that are computed [see Brateley and Schrage (1987) for techniques for assessing the variance of a mean of a correlated sequence of random variables].

4. EXAMPLE: HENRY AARON'S HOME RUN DATA

We illustrate the two models discussed in Section 2 using home run statistics from the career of Henry Aaron. Figure 1 plots the observed home run rates $\{y_i\}$ for the 23 years of Aaron's career. Using the Poisson posterior (3) together with Jeffreys' prior, the posterior mean and standard deviation of the rate λ are given by .0611 and .0022, respectively.

For the Poisson outlier model of Section 2.2, we let the precision hyperparameter $\nu = 10$ and simulated 250 values of the posterior distribution of $(\lambda, \gamma_1, \ldots, \gamma_{23})$ using the "one long chain" method described in Section 3. Figure 2 plots the error bars corresponding to the posterior mean $\pm$ posterior standard deviation for the 23 scale parameters $\{\gamma_i\}$. For Years 1, 18, 20, and 22, note that the error bars lie under the value $\gamma = 1$. Note from (15) that small values of γ_i correspond to values of y_i where the deviance $D(y_i, \lambda)$ is large. Thus, these small values of γ_i indicate that Aaron's performance during these years is inconsistent with his overall performance. Looking at Figure 1, it can be seen that Years 1 and 22 correspond to unusually poor years and Years 18 and 20 to unusually good home run hitting years.

The posterior mean and standard deviation of λ using the outlier model are .0618 and .0027. Note that the standard deviation is 23% higher than the posterior

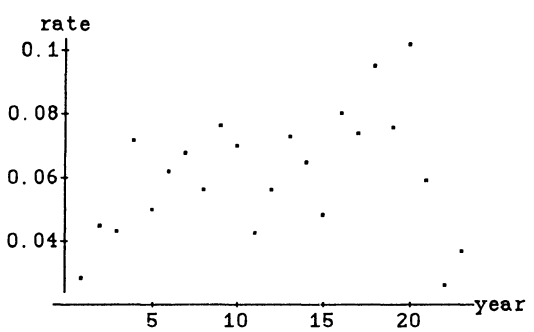

Figure 1. Home Run Rates of Hank Aaron. His seasonal home run statistics are: (HR, AB) = {(13, 468), (27, 602), (26, 609), (44, 615), (30, 601), (39, 629), (40, 590), (34, 603), (45, 592), (44, 631), (24, 570), (32, 570), (44, 603), (39, 600), (29, 606), (44, 547), (38, 516), (47, 495), (34, 449), (40, 392), (20, 340), (12, 465), (10, 271)}.

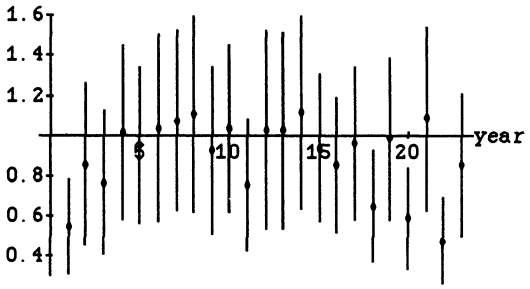

Figure 2. Error Bars for the Posterior Distributions of the Scale Parameters.

standard deviation using the Poisson model. This increase in variability is due to the downweighting of the inconsistent home run years and the uncertainty in the exact location of the $\{\gamma_i\}$.

In Figure 3 error bars for the true rate λ are graphed for all 12 players using the Poisson and Poisson outlier models. The two set of error bars have similar patterns; for example, Ruth stands out as the best slugger in both plots. In closer examination, note that 8 of the 12 hitters have higher rate estimates using the outlier model—these hitters had poor seasons that were downweighted in the outlier model. A second observation is that the posterior standard deviations are generally 20–30% higher using the outlier model. This increase in the size of the standard errors makes the hitters appear more homogeneous in their home run ability.

To investigate the sensitivity of the posterior analysis with respect to the gamma shape parameter ν, Figure 4 plots the posterior means of the scale parameters γ_i for two different simulation runs with $\nu = 4$ and 10. The line $y = x$ is added to the plot for comparison purposes. Note that the posterior means are significantly smaller for $\nu = 4$ for the scale parameters corresponding to outlying years. Thus the shape parameter ν appears to control the size of downweighting of the seasons that correspond to unusually good or poor home run performances.

5. USING A QUADRATIC LOG-LINEAR MODEL

The above model assumes that a player's true home run ability, as measured by the parameter λ, is constant over the span of his career. This model is unrealistic in that it is generally believed that a player's baseball ability grows as he matures, reaches a peak about the mid-

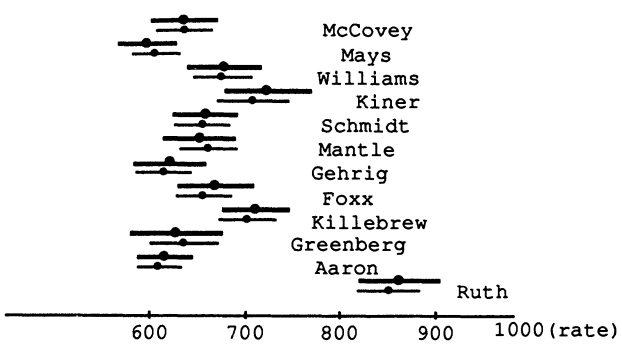

Figure 3. Error Bars for the Home Run Rates for the 12 Players Using Poisson Model (thin bars) and Poisson Outlier Model (thick bars).

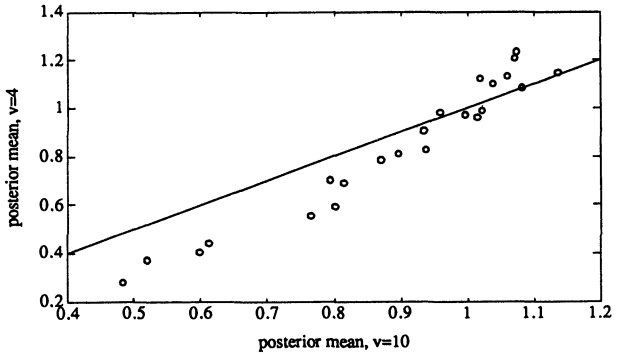

Figure 4. Posterior Means $\{E(\gamma_i|y)\}$ for Two Runs of Simulation Where $\nu = 4$ and $\nu = 10$.

dle of his career, and declines at the end of his career. [See James (1982) for an analysis of baseball players' peak performances.] This belief is substantiated by the location of the posterior distributions of the scale parameters in the constant rate model. Note from Figure 2 that three of the four unusual locations occur during the early and late years in Aaron's career. Thus a more realistic model assumes that the rate for Year i, λ_i, satisfies the log-linear model

$$\log\lambda_i = \beta_0 + \beta_1 i + \beta_2 i^2. \tag{17}$$

In general, suppose $t_i y_i$ is distributed according to the QL density (9) with parameter λ_i, $i = 1, \ldots, p$, where the rates satisfy the log-linear model

$$\log\lambda_i = \mathbf{x}_i^T\boldsymbol{\beta}, \tag{18}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)$ is the vector of unknown regression coefficients and $\mathbf{x}_i^T = (x_{i1}, \ldots, x_{ik})$ is a vector of known constants. If $\boldsymbol{\beta}$, $\gamma_1, \ldots, \gamma_p$ are assumed independent with $\boldsymbol{\beta}$ assigned a noninformative constant prior and γ_i assigned the gamma prior (10), then the joint posterior is given by

$$\pi(\boldsymbol{\beta}, \gamma_1, \ldots, \gamma_p|\mathbf{y}) = \exp\left\{-\frac{1}{2}\sum_{i=1}^{p}\gamma_i[D(y_i, \mathbf{x}_i^T\boldsymbol{\beta}) + \nu]\right\} \times \prod_{i=1}^{p}\gamma_i^{(\nu+1)/2-1} \tag{19}$$

Note that, if the γ_i's are held fixed, then (19) is a generalized linear model likelihood (Nelder and McCullagh 1989) with the addition of scale parameters $\gamma_1, \ldots, \gamma_p$. The mode $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, conditional on $\{\gamma_i\}$, can be found using the usual generalized linear models iterative reweighted least squares algorithm. Then, approximately, $\boldsymbol{\beta}|\gamma_1, \ldots, \gamma_p$ is distributed k-variate multivariate normal with mean $\hat{\boldsymbol{\beta}}$ and variance–covariance matrix $\mathbf{V}$, where $\mathbf{V} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}$,

$$\mathbf{X} = \begin{pmatrix}\mathbf{x}_1^T \\ \cdots \\ \mathbf{x}_p^T\end{pmatrix},$$

and $\mathbf{W}$ is the diagonal matrix with diagonal terms

$t_i \gamma_i \exp\{\mathbf{x}_i^T \hat{\boldsymbol{\beta}}\}$. In addition,

$\gamma_i|\boldsymbol{\beta}, \mathbf{y}, \{\gamma_j, j \neq i\}$ is distributed

$$\text{gamma}\left(\frac{\nu + 1}{2}, \frac{D(y_i, \mathbf{x}_i^T\boldsymbol{\beta}) + \nu}{2}\right). \quad (20)$$

Note that both the multivariate normal and gamma distributions are easy to simulate, so the Gibbs sampler algorithm can be implemented.

For Henry Aaron's home run data, the simulation algorithm was run on the quadratic model (11) with $t = 10$ and $m = 100$. Figure 5 plots error bars for the scale parameters. Comparing this figure with Figure 2, this quadratic model appears to better fit all of the year data. The locations of the posterior densities of γ_i for early and late years are much closer to the value $\gamma = 1$ and the only noticeably poor fit corresponds to the 22nd year. Also, note that the size of the posterior standard deviations of the scale parameters are much higher for this quadratic model, reflecting greater posterior uncertainty about their locations.

Next, consider the posterior density of the fitted rate of the ith year

$$\lambda_i(\boldsymbol{\beta}) = \exp\{\mathbf{x}_i^T\boldsymbol{\beta}\} = \exp\{\beta_0 + \beta_1 i + \beta_2 i^2\}.$$

For these examples, first and second moments of the regression vector $\boldsymbol{\beta}$ were computed using the Gibbs sampler algorithm. If we assume that the posterior distribution of $\boldsymbol{\beta}$ is approximately multivariate normal $(\tilde{\boldsymbol{\beta}}, \boldsymbol{\Sigma})$, where $\tilde{\boldsymbol{\beta}}$, $\boldsymbol{\Sigma}$ are the posterior mean and variance–covariance matrix, then $\lambda_i(\boldsymbol{\beta})$ has mean and variance

$$E[\lambda_i(\boldsymbol{\beta})|\mathbf{y}] = \exp\left\{\mathbf{x}_i^T\tilde{\boldsymbol{\beta}} + \frac{1}{2}\mathbf{x}_i^T\boldsymbol{\Sigma}\mathbf{x}_i\right\}$$

and

$$\text{var}[\lambda_i(\boldsymbol{\beta})|\mathbf{y}]$$

$$= \exp\left\{2\mathbf{x}_i^T\tilde{\boldsymbol{\beta}} + x_i^T\boldsymbol{\Sigma}\mathbf{x}_i\right\}\left[\exp\left\{\mathbf{x}_i^T\boldsymbol{\Sigma}\mathbf{x}_i\right\} - 1\right]. \quad (21)$$

Figure 6 plots the posterior means of the rates as a line on top of the data. In addition, the error limits, mean ± standard deviation, are plotted as dashed lines.

Using this quadratic model, we can distinguish between the peak and career home run performance of a player. We define the peak rate as the maximum value of the fitted rates over the p years; that is,

$$p(\boldsymbol{\beta}) = \max_i \lambda_i(\boldsymbol{\beta}) = \max_i \exp\{\beta_0 + \beta_1 i + \beta_2 i^2\}. \quad (22)$$

To measure the career home run performance, we wish

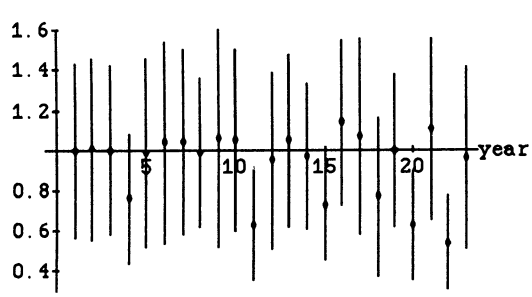

Figure 5. Error Bars of Posterior Densities of Scale Parameters Using Quadratic Model.

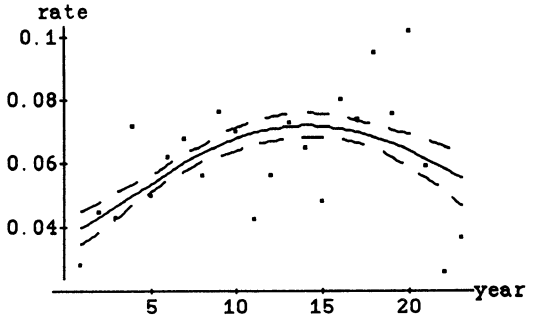

Figure 6. Posterior Means and Error Limits for Fitted Rates Using Quadratic Model.

to accumulate the fitted rates over all years of a player's career. One possible measure is the sum of the fitted rates:

$$\sum_{i=1}^{p} \exp\{\beta_0 + \beta_1 i + \beta_2 i^2\}.$$

However, this measure does not take into account the number of at-bats for different seasons. A part-time player may hit a high rate of home run for many years without accumulating a large number of home runs over his career. Thus we weight each season's rate by the corresponding number of at-bats, obtaining the career measure

$$c(\boldsymbol{\beta}) = \sum_{i=1}^{p} t_i \exp\{\beta_0 + \beta_1 i + \beta_2 i^2\}. \quad (23)$$

If we assume the constant rate Poisson model of Section 2.1 with $\beta_1 = \beta_2 = 0$, this career statistic reduces to the total number of home runs in a career. The measure (23) is an improvement over the total number of home runs in that it accommodates outliers and assumes a more realistic quadratic fit to the home run performance.

Posterior moments of the peak and career home run measures can be computed using the assumption of approximate normality for the regression coefficient β. Equations (21) are used to obtain posterior moments of the peak measure $p(\boldsymbol{\beta}) = \max_i \lambda_i(\boldsymbol{\beta})$. A similar calculation gives the posterior mean of the career measure $c(\boldsymbol{\beta})$:

$$E[c(\boldsymbol{\beta})|\mathbf{y}] = \sum_{i=1}^{p} t_i E[\lambda_i(\boldsymbol{\beta})|\mathbf{y}]$$

$$= \sum_{i=1}^{p} t_i \exp\left\{\mathbf{x}_i^T\tilde{\boldsymbol{\beta}} + \frac{1}{2}\mathbf{x}_i^T\boldsymbol{\Sigma}\mathbf{x}_i\right\}. \quad (24)$$

The posterior variance of $c(\boldsymbol{\beta})$ can be computed by first writing it as

$$\text{var}[c(\boldsymbol{\beta})|\mathbf{y}] = \sum_{i=1}^{p} t_i^2 E[e^{2\mathbf{x}_i^T\boldsymbol{\beta}}|\mathbf{y}]$$

$$+ 2\sum_{i<j} t_i t_j E[e^{2(\mathbf{x}_i^T + \mathbf{x}_j^T)\boldsymbol{\beta}}|\mathbf{y}] - (E[c(\boldsymbol{\beta})|\mathbf{y}])^2, \quad (25)$$

and then computing the expectations in (25) using the formula in (15).

6. SUMMARY

Table 2 presents posterior means and standard deviations of the peak and career measures for the 12 hitters. This table can be compared with the usual home run measures given in Table 1. Several observations can be made in this comparison. First, the career measures for the players in Tables 1 and 2 are similar in size. It is interesting to note that Aaron's career measure increased from 755 to 766, while Ruth's measure decreased from 714 to 706. This suggests that Aaron was a particularly consistent home run hitter during his career and baseball historians generally are in agreement with this conclusion. There exist larger differences between the rate statistics of Table 1 and the peak statistics of Table 2. Mickey Mantle, for example, hit home runs at an overall rate of 6.6%; however, his peak rate was two percentage points higher. It is interesting to note that Mantle is generally considered one of the greatest players at his peak (see James 1988).

One can also use the standard deviations given in this table to compare the performances of various players. Since all of the peak rate standard deviations are approximately equal to .006, the standard deviation of the difference in two players' rates, $p_i(\beta) - pj(\beta)$ is approximately equal to $\sqrt{2(.006)^2} = .0085$. Babe Ruth stands out as the premier home run hitter at his peak, and Kiner, Foxx, Mantle, and Killebrew appear to have a greater peak performance then the bottom group of McCovery, Williams, Aaron, Gehrig, and Mays.

Since this article is written from a Bayesian perspective, it is instructive to compare this methodology with non-Bayesian approaches for handling outliers. Comprehensive treatments of classical model checking for the normal error model are given in the books by Cook and Weisberg (1982) and Belsley, Kuh, and Welsch (1980). Pregibon (1981) describes similar techniques for binomial logistic models and the formula in his article can easily be modified for the Poisson log-linear model considered here.

Let $\{\hat{\lambda}_i\}$ denote the fitted values of $\{\lambda_i\}$ under a particular model. One standard model checking technique plots the components χ^2, $t_i(y_i - \hat{\lambda}_i)/\sqrt{t_i\hat{\lambda}_i}$, or the components of the deviance, $\text{sgn}(y_i - \lambda_i)\sqrt{2t_iy_i \log(y_i/\hat{\lambda}_i)}$, to detect points that are inconsistent with the main body of data. A second diagnostic tool is concerned with the influence of an individual point on the estimated regression coefficients. Pregibon (1981) describes various ways of measuring the change in the fitted regression coefficient when the ith observation is deleted.

Table 2. Ranking of 12 Hitters With Respect to Peak and Career Posterior Means

	Career			Peak	
Hitter	Mean	SD	Hitter	Mean	SD
H. Aaron	766	(25)	B. Ruth	.101	(.008)
B. Ruth	706	(49)	R. Kiner	.087	(.004)
W. Mays	657	(31)	J. Foxx	.086	(.006)
H. Killebrew	581	(21)	M. Mantle	.086	(.006)
M. Schmidt	560	(21)	H. Killebrew	.084	(.005)
M. Mantle	542	(24)	H. Greenberg	.080	(.007)
J. Foxx	537	(30)	M. Schmidt	.079	(.005)
W. McCovey	523	(23)	W. McCovey	.075	(.004)
T. Williams	523	(22)	T. Williams	.075	(.005)
L. Gehrig	497	(28)	H. Aaron	.072	(.004)
R. Kiner	369	(9)	L. Gehrig	.072	(.005)
H. Greenberg	329	(21)	W. Mays	.070	(.004)

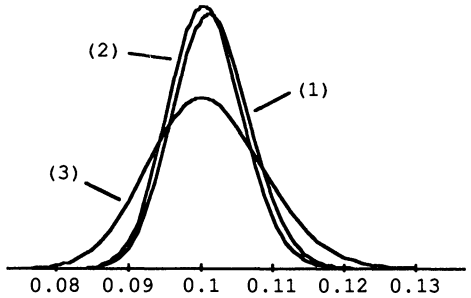

Figure 7. Three Posterior Distributions for the Peak Home Run Rate for Babe Ruth. (1) is the density conditional on $\gamma = 1$; (2) is the density conditional on $\gamma = \tilde{\gamma}$; and (3) is the unconditional density.

Once an aberrant or unusually influential observation is found, one must decide what to do about it. One choice is to delete the observation or give it smaller weight in the analysis. Another choice is to change the model to accommodate the unusual observation.

The above aims of classical model checking can be realized using a Bayesian approach. Plots of the posterior distributions of the individual scale parameters (such as given in Figs. 2 and 4) play a similar role to classical residual plots. Observations that have scale parameter posterior distributions located near $\gamma = 1$ are consistent with the rest of the data. Aberrant points with posterior distributions concentrated on small values have large deviance or chi-square residuals.

Sensitivity of regression estimates with respect to individual observations can be investigated by viewing various conditional posterior distributions. The posterior distribution of $\boldsymbol{\beta}$ conditional on $\gamma_i = 1$ for all i is the analysis with all of the observations given full weight. If $\tilde{\gamma}_i$ denotes the posterior mean of γ_i, then the posterior distribution of $\boldsymbol{\beta}$ conditional on $\gamma_i = \tilde{\gamma}_i$ for all i is the analysis with aberrant observations downweighted. In Figure 6, the posterior densities of Babe Ruth's peak measure $p(\boldsymbol{\beta})$ are plotted conditional on $\gamma_i = 1$ and $\gamma_i = \tilde{\gamma}_i$. Here the two distributions are very similar; the conclusion is that the peak measure is not significantly influenced by any aberrant observations.

The third density in Figure 7 is the unconditional posterior density of $p(\boldsymbol{\beta})$. This distribution accommodates the outliers and, in addition, reflects uncertainty about the presence of outliers. In this example, the standard deviation of the unconditional density is .079, which is approximately 50% larger than the standard deviations of the two conditional posterior densities. Thus, there is a substantive cost in terms of posterior variability in modeling the outliers instead of using a standard Poisson model.

[*Received October 1990. Revised January 1992.*]

REFERENCES

Albert, J. H., and Pepple, P. A. (1989), "A Bayesian Approach to Some Overdispersion Models," *Canadian Journal of Statistics*, 17, 333–344.

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980), *Regression Diagnostics*, New York: John Wiley.

Brateley, P., Fox, B. L., and Schrage, L. E. (1987), *A Guide to Simulation*, New York: Springer-Verlag.

Cook, R. D., and Weisberg, S. (1982), *Residuals and Influence in Regression*, New York: Chapman and Hall.

Gelfand, A. E., Hills, S. E., Racine-Poon, A., and Smith, A. F. M. (1990), "Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling," *Journal of the American Statistical Association*, 85, 972–985.

Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities." *Journal of the American Statistical Association*, 85, 398–409.

Geman, S., and Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.

James, B. (1982), *The Bill James Baseball Abstract*, New York: Ballentine Books.

———. (1988), *The Historical Baseball Abstract*, New York: Villard Books.

Lawless, J. F. (1987), "Negative Binomial and Mixed Poisson Regression," *Canadian Journal of Statistics*, 15, 209–225.

Morris, C. (1983), "Parametric Empirical Bayes Inference: Theory and Applications," *Journal of the American Statistical Association*, 78, 47–54.

Neft, D. S., and Cohen, R. M. (1985), *The Sports Encyclopedia: Baseball*, New York: St. Martins/Marek.

Nelder, J. A., and McCullagh, P. (1989), *Generalized Linear Models*, New York: Chapman and Hall.

Pregibon, D. (1981), "Logistic Regression Diagnostics," *The Annals of Statistics*, 9, 705–724.

Siwoff, S., Hirdt, S., Hirdt, T., and Hirdt, P. (1991), *The 1991 Elias Baseball Analyst*, New York: Fireside.

Thall, P. F. (1988), "Mixed Poisson Likelihood Regression Models for Longitudinal Interval Count Data," *Biometrics*, 44, 197–209.

Thorn, J., and Palmer, P. (1984), *The Hidden Game of Baseball*, New York: Doubleday.

West, M. (1984), "Outlier Models and Prior Distributions in Bayesian Linear Regression," *Journal of the Royal Statistical Society*, Ser. B, 46, 431–439.

——— (1985), "Generalized Linear Models: Scale Parameters, Outlier Accommodation and Prior Distributions," in *Bayesian Statistics 2 (Proceedings of the Second Valencia International Meeting on Bayesian Statistics)*, Amsterdam: North Holland, pp. 531–558.