Journal of Data Science **2**(2004), 61-73

A Two-Stage Bayesian Model for Predicting Winners in Major League Baseball

Tae Young Yang[1] and Tim Swartz[2]
[1]*Myongji University and* [2]*Simon Fraser University*

Abstract: The probability of winning a game in major league baseball depends on various factors relating to team strength including the past performance of the two teams, the batting ability of the two teams and the starting pitchers. These three factors change over time. We combine these factors by adopting contribution parameters, and include a home field advantage variable in forming a two-stage Bayesian model. A Markov chain Monte Carlo algorithm is used to carry out Bayesian inference and to simulate outcomes of future games. We apply the approach to data obtained from the 2001 regular season in major league baseball.

Key words: Major league baseball, Markov chain Monte Carlo, predictive distributions.

1. Introduction

The probability of winning a game in major league baseball (MLB) depends on various factors relating to team strength including the past performance of the two teams, the batting ability of the two teams and the starting pitchers. We define the relative strength of a team over a competing team at a given point in time by combining these three factors into a single measurement. Although other factors relating to team strength may influence the probability of winning, we assume that their effect is minor, and note that the three main factors mentioned above are those that are traditionally considered in the setting of betting odds by bookmakers (McCune 1989). Bookmakers are also aware that the home field advantage significantly influences the probability of winning, and likewise, we include it in our calculations. In this paper, we propose a two-stage Bayesian model based on the relative strength variable and the home field advantage variable to predict the outcomes of games in MLB.

MLB in the United States is divided into two leagues and six divisions. The American League (AL) has three divisions and the National League (NL) has three divisions. Each team plays 162 games in the regular season (April through

Table 1. Summary statistics for the 2001 MLB regular season.

Division	Team	Overall Win %	Home Win%	Away Win %	Team Batting Avg	Team ERA
AL East	NY Yankees	0.59	0.65	0.54	0.267	4.04
	Boston	0.51	0.51	0.51	0.267	4.18
	Toronto	0.49	0.49	0.50	0.263	4.29
	Baltimore	0.39	0.38	0.41	0.248	4.71
	Tampa Bay	0.38	0.46	0.31	0.258	4.96
AL Central	Cleveland	0.56	0.55	0.57	0.278	4.65
	Minnesota	0.53	0.58	0.47	0.272	4.51
	White Sox	0.51	0.57	0.46	0.268	4.55
	Detroit	0.41	0.46	0.36	0.260	5.01
	Kansas City	0.40	0.43	0.37	0.266	4.87
AL West	Seattle	0.72	0.70	0.73	0.288	3.54
	Oakland	0.63	0.65	0.60	0.264	3.59
	Anaheim	0.46	0.48	0.44	0.261	4.20
	Texas	0.45	0.50	0.40	0.275	5.71
NL East	Atlanta	0.54	0.49	0.59	0.260	3.59
	Philadelphia	0.53	0.58	0.48	0.260	4.17
	NY Mets	0.51	0.54	0.47	0.249	4.08
	Florida	0.47	0.58	0.37	0.264	4.32
	Montreal	0.42	0.42	0.42	0.253	4.68
NL Central	Houston	0.57	0.54	0.60	0.272	4.39
	St. Louis	0.57	0.66	0.49	0.270	3.96
	Chicago Cubs	0.54	0.59	0.49	0.261	4.03
	Milwaukee	0.42	0.44	0.40	0.251	4.65
	Cincinnati	0.41	0.33	0.48	0.262	4.78
	Pittsburgh	0.38	0.47	0.30	0.247	5.05
NL West	Arizona	0.57	0.59	0.54	0.267	3.88
	San Francisco	0.56	0.60	0.51	0.266	4.19
	Los Angeles	0.53	0.54	0.52	0.255	4.25
	San Diego	0.49	0.43	0.54	0.252	4.52
	Colorado	0.45	0.51	0.40	0.292	5.29

early October) which does not include the pre-season and the post-season playoffs. Summary statistics for the 2001 MLB regular season appear in Table 1. Overall win percentages are followed by the home and away win percentages, the overall team batting average and the overall team earned run average (ERA).

We observe that most teams win more often at home than on the road, and this suggests that the home field advantage variable may be useful. The 2001 regular season is noteworthy in that Seattle finished with an outstanding 116-46 won-loss record, tying the 1906 Chicago Cubs' MLB record for wins in a season. Seattle also established the AL record for road wins in a season (59) and set a major league record with 29 consecutive road series won or tied. More detailed statistical information on MLB is available from numerous internet websites including www.sportline.com and www.sportsillustrated.com.

The *relative strength* of a team over a competing team at a given point in time is based on three ratios: (1) the ratio of the winning percentages between the two teams, (2) the ratio of the overall team batting averages and (3) the ratio of the ERAs between the two starting pitchers. We note that on the day of a MLB game, all three ratios are typically available. Later we demonstrate that ratio (3) is more important in determining the probability of a win than ratio (1) and ratio (2). The importance of ratio (3) is consistent with our intuition where we believe, for example, that in the 2001 season, the Arizona Diamondbacks have a higher probability of winning when either Randy Johnson or Curt Schilling is pitching. Clearly, it is unreasonable to expect that all three ratios affect the probability of winning in the same way. For this reason, we adopt three unknown contribution parameters that assign relative importance to each of the three ratios. The contribution parameters are assumed constant across all teams and arise from independent prior densities. The home field advantage parameter is assumed to be apriori independent of the contribution parameters. Therefore our model essentially relies on four primary parameters; the three contribution parameters and the home field advantage parameter. This parametrization is much simpler than many other approaches. For example, the Bradley and Terry (1952) model requires at least $n-1$ parameters where a team parameter is defined for each of the n teams and the team parameters sum to a constant. In addition, our relative strength variable is flexible as it readily accommodates other factors, and can be easily modified for various sports such as basketball, football, soccer and hockey.

A two-stage Bayesian model based on the relative strength variable and the home field advantage variable is proposed to predict the outcome of games in MLB. In the first stage, we assume that the probability that a given team wins is a random sample from a beta distribution with parameters based on the relative strength variable and the home field advantage variable. In the second stage, the outcome (win or loss) is a random sample from a Bernoulli distribution with this winning probability. We contrast the two-stage model with a one-stage model where the outcome (win or loss) is a random sample from a Bernoulli(p) distribution whose probability p is a direct function of the relative strength variable and the home field advantage variable. Thus, the two-stage model offers an extra level of variation to account for the difficulty that investigators have encountered when modelling outcomes of MLB baseball games (Kaigh 1995 and McCune 1989). In addition, the two-stage Bayesian structure permits convenient Bayesian inference.

We observe that the posterior distribution arising from the two-stage Bayesian model is of a form that is not readily interpretable. In this context, Markov chain

Monte Carlo (MCMC) algorithms sample variates according to a Markov chain whose stationary distribution is the desired posterior distribution. The goal is to average the sampled variates to estimate posterior quantities of interest. We consider a "Metropolis within Gibbs" MCMC algorithm (see Gilks, Richardson and Spiegelhalter 1996) which is useful when the full conditional densities of the Gibbs sampling algorithm (Geman and Geman 1984) are non-standard. The Gibbs sampling algorithm is perhaps the simplest of the MCMC algorithms. Additional reading on Gibbs sampling in related contexts can be found in Gelfand and Smith (1990), Barry and Hartigan (1993), and Kuo and Yang (1995, 1996, 2000). MCMC sampling is used to simulate the outcomes of future games and then predict the eventual division winners.

Kaigh (1995) considers a simple method of prediction for major league baseball using only the home and away records of the competing teams. Predictions are compared against results from the 1989-1993 MLB regular seasons. It is not evident that the simple predictive model yields a profitable betting strategy. Barry and Hartigan (1993) consider a more complex choice model for prediction in MLB. Like our model, the Barry and Hartigan model is Bayesian, allows for changing team strengths over time and relies on a MCMC implementation. Some distinguishing features of their model include the assumption that team strengths are discrete, the exclusion of team batting averages as covariates and the huge number of parameters required. Perhaps the greatest practical difference, however, is that their model does not consider the effect of starting pitchers. Since starting pitchers may critically affect the outcome of a game, the Barry and Hartigan (1993) model cannot be used for accurate prediction of individual games. Their model is appropriate only for predicting a large number of games (e.g. the remainder of a season) where the effect of individual pitchers is averaged over the large number of games. Additional reading on baseball and prediction can be found in Bennett (1998), James, Albert and Stern (1997), and Barry and Hartigan (1994).

Section 2 provides a complete specification of the full two-stage Bayesian model and indicates sub-models that may be appropriate. In Section 3, we adopt a Markov chain Monte Carlo approach for Bayesian inference. The full conditional distributions are derived and sampling is carried out using the Metropolis within Gibbs algorithm. Section 4 looks at the problem of prediction with specific emphasis on the prediction of results for the remainder of the season. In Section 5, simulation results from the models are studied in comparison with the actual final results from the 2001 MLB regular season. Some concluding remarks are then provided in Section 6.

2. The Model

Suppose that the T games in a MLB regular season are indexed in time according to $s = 1, \ldots, T$. We calculate covariates $(\alpha_s, \beta_s, \gamma_s)$ immediately prior to time s where

- α_s denotes the ratio of the winning percentages between the two teams

- β_s denotes the ratio of the overall team batting averages
- γ_s denotes the ratio of the starting pitchers' ERAs

Without loss of generality, we let the numerator in the first two ratios correspond to the home team, and we let the denominator in the third ratio correspond to the home team.

As pointed out by a referee, there is variability in the covariates and the variability is decreasing in time s. In particular, early in the season, some pitchers may have extraordinarily high or low ERAs. For this reason, we truncate γ_s according to $1/\gamma_0 \leq \gamma_s \leq \gamma_0$ for prescribed γ_0. For the MLB data considered in this paper, we let $\gamma_0 = 5$. From a practical point of view, we consider only games corresponding to $s \geq t_0$ for some moderate value t_0, to allow the covariates to stabilize. Typically, with MLB data, one would not expect extreme variation in the covariates α_s and β_s for $s \geq t_0$.

It is unreasonable to expect that all three ratios affect the probability of winning in the same way. We therefore adopt unknown contribution parameters $r = (r_1, r_2, r_3)$ where we define the relative strength of the home team over the visiting team at time s as

$$\lambda_s = \alpha_s^{r_1} * \beta_s^{r_2} * \gamma_s^{r_3}.$$

We assume that the contribution parameters are constant across all MLB teams and arise from independent prior distributions with $r_i \sim \text{uniform}(0, a_i)$ where a_i is prescribed, $i = 1, 2, 3$. The value r_i close to 0 implies that the corresponding ratio has little effect on the relative strength. For the MLB data considered in this paper, we let $a_i = 2$ for $i = 1, 2, 3$. We view these as subjective priors where we use our knowledge of MLB to impose effects based on the current winning percentage, batting and pitching. Note that in the context of the models described below, $a_i = 2$ is a realistic upper bound for the value of r_i, $i = 1, 2, 3$.

The relative strength of the visiting team over the home team at time s is therefore given by the reciprocal $1/\lambda_s = (1/\alpha_s)^{r_1} (1/\beta_s)^{r_2} (1/\gamma_s)^{r_3}$. When a team's relative strength is larger (smaller) than 1, this implies that it is the stronger (weaker) team at time s. We note that the relative strength variable λ_s is flexible as it readily accommodates other factors, and can be easily modified for various sports.

We also consider a home field advantage variable δ which is assumed constant across all MLB teams and is assumed independent of the contribution parameters. The visiting field effect is then defined as $1/\delta$ where we interpret $\delta > 1$ as a home field advantage. We assign a uniform(δ_0, δ_1) prior distribution for δ where δ_0 and δ_1 are prescribed. For the MLB data considered in this paper, we let $\delta_0 = 0$ and $\delta_1 = 2$. The apriori belief $E(\delta) = 1$ therefore implies that the home field has no effect.

Having defined the relative strength λ_s and the home field advantage δ, we now propose three predictive models in increasing levels of complexity. Let the random variable X_s equal 1 (0) if the home team wins (loses) in the sth game. We are interested in the probability distribution of X_s. For convenience, we define $p_s = \text{Prob}(X_s = 1)$.

$$\text{Model 1:}\quad \text{Prob}(X_s) = p_s^{X_s}(1-p_s)^{1-X_s} \quad \text{and} \quad p_s = \frac{\lambda_s\delta}{1+\lambda_s\delta}$$

Model 1 is a simple one-stage model where X_s is a Bernoulli random variable with expected value p_s. One of the features of the parametrization is that the probability of the visiting team winning is given by $\left(\frac{1}{\lambda_s\delta}\right) / \left(1 + \frac{1}{\lambda_s\delta}\right) = \frac{1}{1+\lambda_s\delta} = 1 - p_s$.

$$\text{Model 2:}\quad \text{Prob}(X_s) = p_s^{X_s}(1-p_s)^{1-X_s} \quad \text{and} \quad p_s \sim \text{beta}(\lambda_s\delta, 1)$$

Model 2 is a two-stage model where the first stage uses the same Bernoulli distribution as in Model 1 and the second stage imposes a beta prior distribution. Note that conditional on p_s, the expected value of X_s is again p_s. Model 2 offers an additional level of variation beyond Model 1 to account for the difficulty that investigators have encountered when modelling the outcomes of MLB games.

$$\text{Model 3:}\quad \text{Prob}(X_s) = p_s^{X_s}(1-p_s)^{1-X_s} \quad \text{and} \quad p_s \sim \text{beta}(m\lambda_s\delta, m)$$

Model 3 generalizes Model 2 by introducing the parameter $m > 0$. When m is equal to 1, then Model 3 reduces to Model 2. When $m \to \infty$, then Model 3 reduces to Model 1. Thus Model 1 and Model 2 can be viewed as sub-models, and we can assess their suitability according to the posterior distribution of m in Model 3. Note that Model 3 implies $\text{E}(p_s \mid m) = \lambda_s\delta/(1+\lambda_s\delta)$ as in Model 1 and in Model 2, and $\text{Var}(p_s \mid m) = \lambda_s\delta/(\lambda_s\delta+1)^2(m\lambda_s\delta+m+1)$. Therefore we can maintain the same mean structure, yet the probability distribution can be made more or less diffuse according to the value of m. We choose an exponential prior distribution with mean m_0 for m. In this paper, we let $m_0 = 10$.

3. Bayesian Inference

Consider inference for the primary parameters r, δ, m and the p_s in the full two-stage Bayesian model (Model 3). We refer to the fixed parameters a_1, a_2, a_3, δ_0, δ_1 and m_0 as hyperparameters of the model.

In the 2001 MLB regular season, each of the 30 teams is scheduled to play 162 games leading to a total of $T = 15(162) = 2430$ games. Again, the games are indexed in time according to $s = 1, \ldots, t, \ldots, T$ where t is the current time. We denote the covariate triple

$$D_s = (\alpha_s, \beta_s, \gamma_s) \quad \text{for} \quad s = t_0, \ldots, t-1$$

where the components of D_s are the three ratios defined in Section 2. Considering games from only time t_0 onwards, the posterior density of $p_{t_0}, \ldots, p_{t-1}$, m, δ and r is therefore proportional to the likelihood times prior and is given by

$$\begin{aligned} \pi(\cdot \mid \cdot) &= \pi(p_{t_0}, \ldots, p_{t-1}, m, \delta, r \mid x_{t_0}, \ldots, x_{t-1}, D_{t_0}, \ldots, D_{t-1}) \\ &\propto \left(\prod_{s=t_0}^{t-1} B(m\lambda_s\delta, m)\, p_s^{x_s+m\lambda_s\delta-1}(1-p_s)^{m-x_s}\right) e^{-m/m_0} \end{aligned} \tag{3.1}$$

where $B(a,b) = \Gamma(a+b)/(\Gamma(a)\Gamma(b))$ is the norming constant of the beta(a,b) distribution, $m > 0$, $\delta_0 \le \delta \le \delta_1$, $a_i \le r_i \le b_i$ for $i = 1,2,3$ and $0 < p_s < 1$ for $s = t_0, \ldots, t-1$.

For inference we use the empirical measure of the samples generated by the Gibbs sampler. From (1), the Gibbs sampler requires iterative variate generation from the following four full conditional densities:

$$\begin{array}{lcl}
[p_s \mid \cdot] & \sim & \text{beta}\,(x_s + m\lambda_s\delta,\ m - x_s + 1) \quad s = t_0, \ldots, t-1 \\
[m \mid \cdot] & \propto & e^{-m/m_0} \prod_{s=t_0}^{t-1} B(m\lambda_s\delta, m)\, p_s^{m\lambda_s\delta}\, (1-p_s)^m \\
[\delta \mid \cdot] & \propto & \prod_{s=t_0}^{t-1} \frac{\Gamma(m\lambda_s\delta + m)}{\Gamma(m\lambda_s\delta)} p_s^{m\lambda_s\delta} \\
[r_i \mid \cdot] & \propto & \prod_{s=t_0}^{t-1} \frac{\Gamma(m\lambda_s\delta + m)}{\Gamma(m\lambda_s\delta)} p_s^{m\lambda_s\delta} \quad i = 1,2,3
\end{array}$$

The generation of the p_s variates is straightforward as almost every statistical package has a built-in beta generator. The generation of m, δ, r_1, r_2 and r_3 from their respective full conditional densities is less obvious as their densities have non-standard forms. For these variates, we use a Metropolis within Gibbs step (Gilks, Richardson and Spiegelhalter 1996) where in each case we use the corresponding prior distribution as the proposal distribution. Note that this is a practical advantage in avoiding improper priors.

We now estimate the probability that the home team wins at time t. Denote the data at time t as

$$\text{data} = (x_{t_0}, \ldots, x_{t-1}, D_{t_0}, \ldots, D_{t-1}).$$

Then the predictive density at time t is given by

$$\begin{aligned} P(x_t \mid \text{data}) \quad & \propto \int p_t^{x_t}(1-p_t)^{1-x_t} B(m\lambda_t\delta,\ m)\, p_t^{m\lambda_t\delta - 1}(1-p_t)^{m-1} \\ & \qquad \times \pi(m, \delta, r \mid \text{data})\, dp_t\, dm\, d\delta\, dr \end{aligned} \tag{3.2}$$

where $\pi(m, \delta, r \mid \text{data})$ is the marginal posterior density obtained from (3.1). It is not practical to integrate according to (3.2). Instead, the integration is carried out using the MCMC algorithm for Model 3. Retaining the values (m, δ, r) obtained in an iteration of the MCMC algorithm, generate $p_t \sim \text{beta}(m\lambda_t\delta, m)$ and then generate X_t from a Bernoulli(p_t) distribution. The X_t are now a sample from the predictive density in (3.2) and can be averaged to estimate the predictive probability that the home team wins. Noting that $E(X_t|\text{data}) = E(P_t|\text{data})$, a simpler and more efficient estimate of the predictive probability that the home team wins can be obtained by averaging the $\lambda_t\delta/(\lambda_t\delta + 1)$ values. This technique is sometimes referred to as Rao-Blackwellization.

4. Prediction

Besides making inferences on the unknown parameters, we are also interested in predicting the outcome of games for the remainder of the season. We remark that the MLB schedule is determined well in advance of the season, and therefore we always know who plays whom at a given point in time.

Recall that t is the current time in the season. The predictive density of $X_t, \ldots, X_T$ can be obtained via

$$\begin{aligned} & P(X_t, \ldots, X_T \mid X_{t_0}, \ldots, X_{t-1}, D_{t_0}, \ldots, D_{t-1}) \\ = \; & \textstyle\prod_{s=t}^{T} P(X_s \mid X_{t_0}, \ldots, X_{s-1}, D_{t_0}, \ldots, D_{s-1}) \end{aligned}$$

where the terms within the product are given in (3.2).

Our Markov chain algorithm can also be used for estimating results for the remainder of the season. After the chain has stabilized, we retain the values (m, δ, r) from a given iteration. We generate $p_t \sim \text{beta}(m\lambda_t\delta, m)$ and then generate $X_t \sim \text{Bernoulli}(p_t)$. Next we generate $p_{t+1} \sim \text{beta}(m\lambda_{t+1}\delta, m)$ followed by $X_{t+1} \sim \text{Bernoulli}(p_{t+1})$. We continue the process obtaining a single sequence $X_t, \ldots, X_T$. For each team, we combine the number of wins at the current time in the season with the number of wins during the simulated season $t, \ldots, T$ to obtain the number of wins over the full season. The entire process is repeated over multiple simulated seasons to derive estimates for the total number of wins. A practical difficulty concerns the updating of $D_s = (\alpha_s, \beta_s, \gamma_s)$ used in the construction of λ_s during the simulated part of the season $s = t, \ldots, T$. The win ratios α_s change for a team as they win and lose games in the simulated part of the season. Although α_s is readily changed according to $X_{t_0}, \ldots, X_{s-1}$, the ratios for team batting β_s and starting pitchers' ERAs γ_s are unavailable for games well off into the future. As a compromise, we fix $\beta_s = \beta_t$ and set γ_s equal to the team average at time t, $s = t, \ldots, T$.

5. Analysis of 2001 MLB Data

Collecting relevant information on the entire 2001 MLB regular season is a formidable task. As such, we considered only games played on each of twelve dates, beginning April 15 and spread nearly evenly across the season. In these games, we observe 106 home team wins out of a total of 179 games giving a home field winning proportion of 0.59. This sample proportion is a little larger than historical values for MLB (Berry 2001). Table 2 lists the sample means and standard errors for α, β, γ and the home field winning variable X. We observe that the sample means of α, β and γ are what we expect, e.g. nearly centered about 1. Note that β does not vary greatly about 1.0; this is an early suggestion that the batting variable may not have a significant role in prediction.

We first consider practical convergence of the Markov chain using the CODA software (Best, Cowles and Vines 1995). CODA is a set of S-Plus functions that gives graphical summaries and diagnostics of convergence from MCMC output. The results from CODA suggest that practical convergence is achieved after 1,000 iterations. The following results are based on 1000 iterations and 100 replications after 1000 burn-in iterations.

Table 3 provides posterior means and posterior standard deviations of the primary parameters r, δ and m. The posterior distributions of the r variables are

Table 2: Sample means and standard errors of α, β, γ and X.

	α	β	γ	X
Mean	1.07	0.99	1.15	0.59
Standard Error	0.03	0.01	0.06	0.04

somewhat flat confirming the difficulty in determining the exact contribution of the winning percentage, batting and pitching to the prediction problem. It also points to the need for larger data sets in future analyses.

Table 3: Priors, posterior means and posterior standard deviations of the primary parameters.

Parameter	Prior	Posterior Mean	Posterior S.D.
r_1	uniform$(0, 2)$	1.23	0.89
r_2	uniform$(0, 2)$	0.73	0.88
r_3	uniform$(0, 2)$	1.17	0.91
δ	uniform$(0, 2)$	1.58	0.21
m	exponential(10)	5.23	3.19

To interpret the magnitude of the home field advantage, consider two teams with equal relative strength $\lambda_s = 1$. In this case, the posterior mean of δ given by 1.58 implies that the home team wins with expected probability 0.61, which is close to the sample mean of X given by 0.59. The posterior mean $m = 5.23$ is less than the prior mean 10.0 and is in the direction of Model 2 where the value of m is 1. This provides some support for Model 2 although we rely on a model choice criterion based on prediction.

We now turn to the prediction problem where the remainder of the 2001 MLB season is simulated from different points in time. We note that there are very few MLB games that are rained out and not rescheduled. Therefore almost all of the 30 teams completed the full 162 games. We compare the predictive ability of the three models given in Section 2. The criterion used is the sum of squared differences between the expected predictive win probabilities and the final season winning percentages over all 30 MLB teams. The chosen current time points are May 30, June 30, July 30 and August 30 in the regular season. The results are shown in Table 4 where the last column indicates the squared difference criterion between the current winning percentages and the final season winning percentages. The last column therefore corresponds to simple extrapolation of a team's current performance to the end of the season. Naturally, we expect final season predictions to improve as the season progresses and this is the case in Table 4. We observe that the predictive ability of all three models is superior to simple extrapolation. Given the aforementioned difficulty of prediction in MLB, we stress that this is a promising result and is indicative of model adequacy. Amongst the three models, it seems that Model 3 may be overfitting although all three models are comparable in their predictive performance. We prefer Model 2 as it is slightly better than Model 3 in prediction and allows more variation than Model 1 without requiring additional parameters. The results from Table 3 also suggest that we might prefer Model 2 over Model 1.

Table 4: Squared error difference criterion for comparing the three models of Section 2 and the simple extrapolation model.

Predictive Date	Model 1	Model 2	Model 3	Extrapolation
May 30	0.089	0.089	0.097	0.154
June 30	0.075	0.075	0.083	0.089
July 30	0.037	0.037	0.039	0.041
August 30	0.011	0.011	0.011	0.012

In Table 5, we provide the prediction results for each of the 30 MLB teams from two points in time (May 30, July 30) based on Model 2. We note that 22 of the 30 predictions from July 30 are closer to the final winning percentages than the current winning percentages recorded on July 30. We also note that the model tends to exhibit a regression towards the mean effect. For example, using results to three decimal places, all six division leaders on July 30 have a predicted winning percentage that is less than their actual winning proportion on July 30.

It is also interesting to compare the predictive impact of covariates α_s, β_s and γ_s. The seven types of covariate combinations under Model 2 are given by

$$\begin{aligned} &S_1 : \lambda_s = \alpha_s^{r_1}, \quad S_2 : \lambda_s = \beta_s^{r_2}, \quad S_3 : \lambda_s = \gamma_s^{r_3}, \quad S_4 : \lambda_s = \alpha_s^{r_1}\beta_s^{r_2}, \\ &S_5 : \lambda_s = \alpha_s^{r_1}\gamma_s^{r_3}, \quad S_6 : \lambda_s = \beta_s^{r_2}\gamma_s^{r_3}, \quad S_7 : \lambda_s = \alpha_s^{r_1}\beta_s^{r_2}\gamma_s^{r_3}. \end{aligned} \tag{5.1}$$

The relative strengths of S_1, S_2 and S_3 consist of only one effect. The relative strengths of S_4, S_5 and S_6 consist of two effects and the relative strength of S_7 consists of all three effects. Table 6 provides the sum of squared differences between the expected predictive win probabilities and the final season winning percentages over all 30 MLB teams based on S_1 to S_7. For comparison purposes, the criterion is also reported for the simple extrapolation model. The results are not definitive but are in accord with prevailing wisdom that suggests that starting pitching is a key determinant in prediction.

Table 5. Predicted probabilities from different points in time using Model 2.

Team	Final Win Percentage	May 30 (Actual , Predicted)	July 30 (Actual , Predicted)
NY Yankees	0.59	0.58 , 0.59	0.61 , 0.60
Boston	0.50	0.55 , 0.57	0.58 , 0.58
Toronto	0.49	0.51 , 0.50	0.45 , 0.47
Baltimore	0.39	0.48 , 0.48	0.42 , 0.42
Tampa Bay	0.38	0.29 , 0.30	0.32 , 0.32
Cleveland	0.56	0.65 , 0.57	0.58 , 0.55
Minnesota	0.52	0.67 , 0.61	0.57 , 0.58
White Sox	0.51	0.38 , 0.44	0.50 , 0.51
Detroit	0.40	0.44 , 0.43	0.44 , 0.43
Kansas City	0.40	0.35 , 0.39	0.39 , 0.41
Seattle	0.71	0.76 , 0.68	0.72 , 0.70
Oakland	0.63	0.48 , 0.53	0.53 , 0.55
Anaheim	0.46	0.46 , 0.50	0.50 , 0.51
Texas	0.45	0.35 , 0.33	0.45 , 0.42
Atlanta	0.54	0.51 , 0.55	0.57 , 0.57
Philadelphia	0.53	0.64 , 0.60	0.54 , 0.54
NY Mets	0.50	0.42 , 0.45	0.46 , 0.47
Florida	0.46	0.46 , 0.49	0.50 , 0.50
Montreal	0.42	0.38 , 0.40	0.42 , 0.43
Houston	0.57	0.49 , 0.49	0.54 , 0.53
St. Louis	0.57	0.58 , 0.57	0.50 , 0.52
Chicago Cubs	0.54	0.60 , 0.59	0.59 , 0.59
Milwaukee	0.42	0.52 , 0.52	0.44 , 0.45
Cincinnati	0.40	0.39 , 0.40	0.39 , 0.40
Pittsburgh	0.38	0.34 , 0.35	0.39 , 0.39
Arizona	0.56	0.57 , 0.57	0.56 , 0.57
San Francisco	0.55	0.50 , 0.49	0.54 , 0.53
Los Angeles	0.53	0.55 , 0.54	0.58 , 0.56
San Diego	0.48	0.52 , 0.50	0.49 , 0.49
Colorado	0.45	0.50 , 0.44	0.42 , 0.41

6. Concluding Remarks

We have seen that the proposed two-stage Bayesian model is effective in predicting division winners in MLB given results partway through the season. The model is simple, easily incorporates other factors, can be extended to other sports and is amenable to a MCMC implementation.

What we have not done in this paper is compare the predictive probabilities of winning with those obtained from the odds posted by bookmakers. Bookmakers use statistical information and intuition in determining their odds. When our predictive probabilities differ from those of the bookmaker by a prescribed margin, this triggers a wagering situation. The natural question then is whether this betting strategy yields income sufficient to overcome the bookmaker's vigorish. This is a topic of future research. More information on aspects of sports gambling can be found in Insley, Mok and Swartz (2003).

Table 6: Squared error difference criterion for assessing covariate combinations as described in (5.1). The simple extrapolation model is also included.

Predictive Date	S_1	S_2	S_3	S_4	S_5	S_6	S_7	Extra-polation
May 30	0.104	0.101	0.075	0.091	0.089	0.071	0.089	0.154
June 30	0.083	0.081	0.067	0.074	0.074	0.066	0.075	0.089
July 30	0.045	0.441	0.039	0.036	0.037	0.038	0.037	0.041
August 30	0.015	0.015	0.013	0.011	0.011	0.013	0.011	0.012

Acknowledgement

Yang's work was partially supported by a grant from the Korea Science and Engineering Foundation. Swartz's work was partially supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

References

Barry, D. and Hartigan, J. A. (1993). Choice models for predicting divisional winners in major league baseball. *Journal of the American Statistical Association,* **88**, 766-774.

Barry, D. and Hartigan, J. A. (1994). Change points in 0-1 sequences, with an application to predicting divisional winners in major league baseball. *Journal of Applied Statistical Science* **1**, 323-336.

Bennett, J.M. (1998). Baseball. In *Statistics in Sport* (Edited by J .M. Bennett). Arnold Applications of Statistics Series, Arnold Publishing, 25-64.

Berry, S. (2001). A Statistician reads the sports pages. *Chance* **14**(1), 52-57.

Best, N. G., Cowles, M. K. and Vines, S. K. (1995). *CODA: Convergence Diagnosis and Output Analysis Software for Gibbs Sampling Output, Version 0.3.* MRC Biostatistics Unit, Cambridge.

Bradley, R. A. and Terry, M. E. (1952). Risk analysis of incomplete block designs I: The method of paired comparisons. *Biometrika* **39**, 324-345.

Gelfand, A. and Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398-409.

Gelman, A. E. and Rubin, D. (1992). Inference from iterative simulation using multiple Ssquences. *Statistical Science* **7**, 457-472.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transaction on Pattern Analysis and Machine Intelligence* **6**, 721-741.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (editors) (1996). *Markov Chain Monte Carlo in Practice*. Chapman and Hall.

Insley, R., Mok, L. and Swartz, T. B. (2003). Some practical results related to sports gambling. To appear in *The Australian and New Zeland Jpurnal of Statistics*.

James, B., Albert, J. and Stern, H.S. (1993). Answering questions about baseball using statistics. *Chance*, 6(2). 17-22,30.

Kaigh, W.D. (1995). Forecasting baseball games. *Chance* 8(2). 33-37.

Kuo, L. and Yang, T. (1995). Bayesian computation for software reliability. *Journal of Computational and Graphical Statistics* **4**, 1-18.

Kuo, L. and Yang, T. (1996). Bayesian computation for nonhomogeneous Poisson processes in software reliability. *Journal of the American Statistical Association* **91**, 763-773.

Kuo, L. and Yang, T. (2000). Bayesian reliability modeling for masked system lifetime data. *Statistics and Probability Letters* **47**, 229-241.

McCune, B. (1989). *Education of a Sports Bettor*. McCune Sports Investments.

Received November 14, 2002; accepted March 11, 2003

Tae Young Yang
Department of Mathematics
Myongji University
Kyunggi, Korea
tyang@wh.myongji.ac.kr

Tim Swartz
Department of Statistics and Actuarial Science
Simon Fraser University
Burnaby, British Columbia
Canada V5A1S6
tim@stat.sfu.ca