Research article

Eric A. E. Gerber* and Bruce A. Craig

A mixed effects multinomial logistic-normal model for forecasting baseball performance

<https://doi.org/10.1515/jqas-2020-0007>
Received January 23, 2020; accepted November 29, 2020;
published online January 6, 2021

Abstract: Prediction of player performance is a key component in the construction of baseball team rosters. As a result, most prediction models are the proprietary property of team or industrial sports entities, and little is known about them. Of those models that have been published, the main focus has been to separately model each outcome with nearly no emphasis on uncertainty quantification. This research introduces a joint modeling approach to predict seasonal plate appearance outcome vectors using a mixed-effects multinomial logistic-normal model. This model accounts for positive and negative correlations between outcomes, both across and within player seasons, and provides a joint posterior predictive outcome distribution from which uncertainty can be quantified. It is applied to the important, yet unaddressed, problem of predicting performance for players moving between the Japanese (NPB) and American (MLB) major leagues.

Keywords: Bayesian hierarchical modeling; Japanese baseball; sports prediction; uncertainty assessment.

1 Introduction

Baseball is one of the most attended spectator sports in the world. Both Major League Baseball (MLB) of the United States and Nippon Professional Baseball (NPB) of Japan are the most attended sporting leagues in their respective countries. Since the reorganization of the NPB into its current form in 1950, nearly every year has had at least one player transition between these two leagues.

There is inherent risk involved in transitioning between the two leagues for both player and team. MLB players

transitioning to Japan tend to cost NPB teams more in salary than Japanese players, and are expected to immediately fill a role as a key contributor. NPB players moving to MLB are more often among the strongest in Japanese baseball, yet are coming from a league most consider weaker than MLB. They also can cost MLB teams a non-trivial amount simply for the rights to negotiate with them.

In both directions of transition, it is important for the team management to determine the worth and level of the added investment. Thus it is imperative to have reliable tools for predicting a player's performance when moving between the United States and Japan. Baseball operations departments have their own methods for predicting performance and access to proprietary data sets that track not only the traditional baseball statistics but also advanced measures, such as pitch spin, batting launch angle, or spatial fielding metrics. To this point, the question of predicting baseball player's performance when moving between MLB and NPB has not been addressed in a public setting. Nor has there been appropriate emphasis on describing uncertainty about predictions in forecasting models for baseball players.

This research develops methodology that will serve, among other applications, to predict performance of baseball players moving between MLB and NPB. We introduce the mixed-effects multinomial logistic-normal model for longitudinal data, followed by tools for assessing model fit, making predictions and assessing uncertainty about those predictions. The methodology is applied to address these NPB/MLB transitions using all players (pre-2015) who have played in both leagues. Predictions are compared to the standard multinomial logistic model, and also to the Marcells, the most established publicly accessible prediction methodology. The multinomial logistic-normal's predictions outperform both of these benchmarks.

2 Background

2.1 Literature review

Most published work in sports prediction focuses on univariate, binary on-field performance metrics. One of the

*Corresponding author: Eric A. E. Gerber, Department of Mathematics, California State University Bakersfield, Science III, Student Way, 93311, Bakersfield, CA, USA, E-mail: egerber1@csb.edu
Bruce A. Craig, Department of Statistics, Purdue University, West Lafayette, IN, USA, E-mail: bacraig@purdue.edu

earliest examples comes from Lindsey (1959), who utilizes proportion tests to predict a baseball player's batting average based on the handedness of the batter-pitcher match-up. Subsequently, in a seminal paper, Efron and Morris (1975) use Stein's estimator via empirical Bayes to predict batting averages for baseball players for the remainder of the 1970 season. Additional works include Bennett and Flueck (1983), Albright (1993), Albert (1994), and Rosner, Mosteller, and Youtz (1996).

More recently, there has been a broader and more rich tapestry of statistical models applied to the problem of prediction in baseball and other sports, including a greater emphasis on hierarchical and non-parametric models. Berry, Reese, and Larkey (1999) use a hierarchical model to compare home run abilities across eras. They applied the same model, separately, to batting average. Chandler and Stevens (2012) use random forests to project future success of minor-league baseball players, where their metric for success was simply the binary outcome of playing in greater than 320 Major League games or not. In a non-baseball application, Wickramasinghe (2014) focuses on predicting average runs for batsmen in cricket via a hierarchical linear mixed model.

Within this realm of published research, there has been little focus on developing models that seek to predict the entire vector of plate appearance outcomes. The three most notable papers to tackle this issue are by Null (2009), Albert (2016) and Powers, Hastie, and Tibshirani (2018). Null (2009) defines 14 distinct outcomes for a plate appearance and assumes each batter has some "true" probability vector for experiencing these outcomes. Null's work also serves as an introduction to the class of nested Dirichlet models, used as an alternative to the Dirichlet or Generalized Dirichlet for modeling multinomial data.

Albert (2016) develops a Bayesian random effects model for estimating component rates of batting outcomes in order to better predict batting average. Albert's work treats the outcomes as multinomial in nature, but factorizes the multinomial likelihood into a product of conditional binomial likelihoods, and assigns independent prior distributions to the probabilities of each outcome. Between Null (2009) and Albert (2016), only Null attempts to include covariates, using a fixed effects regression model to capture an age effect.

Powers, Hastie, and Tibshirani (2018) seek to estimate multinomial probabilities while accounting for relative strength of the batter and pitcher. They focus on individual plate appearances, utilizing play-by-play data. The nuclear penalized multinomial regression approach they develop, similar to Null (2009) and Albert (2016), defines a nesting

structure of batting outcomes which they leverage to allow for positive associations among outcome probabilities and produce better estimates, though they find the key benefit in their model is in the ease of interpretation.

Outside the realm of academic journals, there is an ample and active sector of the Internet where baseball projection systems are implemented. Most projections from these models are published each season on the Internet. Unfortunately, only one of the most common projection systems is completely open source, with all others maintaining some proprietary ownership of the methodology. Not only that, but there is a dearth of uncertainty quantification associated with the resulting projections, which would allow for comparing the "floor" and "ceiling" of different players.

One of the only systems with some level of uncertainty assessment is the *Marcells* (Tango 2004). The *Marcells* is also the only well established, completely open source projection system, with the model details available online (Tango 2004). In brief, the *Marcells* consists of regressing a weighted average of their last three seasons to the mean. Details, including the adjustment for age, are described in Appendix C. In terms of uncertainty, the *Marcells* provides a "reliability score" for each prediction; simply the proportion of the prediction which comes from player data as opposed to the regression to the mean. Players with no data, including minor league players and players coming from Japan, will be predicted at the mean and have a reliability score of 0 (Tango 2012).

Because of the lack of open source methods as well as a lack of uncertainty quantification, we think adding a model to the lexicon of published sports prediction is a useful contribution, especially considering the heretofore unexplored application.

2.2 Data description

Our population of interest in this study is the set of players to play in both NPB and MLB. The data used in this project come from two main sources. Nippon Professional Baseball player-season data were provided by Data Stadium, Inc. Major League Baseball player-season data for relevant players, as well as their corresponding covariates, were taken from the Lahman database available in R (Friendly 2015). Some data augmentation was necessary to collate the two data sources. Some artifacts of the different data sources, such as slight differences in height and weight for the same players, still remain. We also consider players switching between leagues in the middle of a season as two seasons of data.

The data consist of all players who have played in both NPB and MLB before 2015. For the batting data, this includes 605 players with 4991 player-seasons. Of the 605 players, 50 made their initial transition from NPB to MLB, while 555 made the reverse transition. We limit our data to those players making the switch for two main reasons. First, we do not have easy access to all seasons of NPB data. Second, considering players moving from NPB to MLB tend to be among the best in NPB, there is an argument to be made that the players do not belong in the same population. A similar argument could be made that MLB players transitioning to NPB do not belong in the same population as all MLB players. A potential adjustment to the model, which could allow for inclusion of all players from both leagues, given the data, is suggested in the conclusion.

While it is possible to fit a pitching outcome model using this approach (i.e., replace at bats with batters faced), we focus only on modeling hitter outcomes here, though we do include batting outcomes of pitchers and include pitcher as a positional variable.

Choosing the granularity of the count vectors depends on the nature of the data collected and the goals of the research. We consider the following 10 outcomes: base hits (Single, Double, Triple, Home Run), free passes (Base on Balls, Hit-by-Pitch), sacrifices (Sacrifice Bunt, Sacrifice Fly), and outs (Strikeout, Out-in-Play). These give enough information to describe a player's offensive performance, while being sufficient for calculating summary statistics, such as batting average $\left(\frac{H}{AB}\right)$ or on-base plus slugging percentage (OPS) $\left(\frac{H+BB+HBP+TB}{AB}\right)$. The individual base hit categories allow us to calculate total bases (TB) which is used in the calculation of OPS. The two free pass categories are likewise used in calculating OPS, and while the sacrifice categories are not, at-bats (AB) are made up of base hits and outs, and thus distinguishing sacrifices becomes necessary. In developing and testing the methodology, a smaller three outcome model was also used, some results from which are discussed in Appendix B.

Among the covariates available, those chosen were based on what we considered to best explain variability in performance. Height, weight, age, and batting handedness all could reasonably impact the performance of certain types of players; heavier players, for example, will be less likely to be slap hitters looking to take advantage of their speed. A quadratic age term was included, as there have been some studies which suggest the career trajectories of players' performance are quadratic in nature (Albert 2002; Fair 2008). Berry, Reese, and Larkey (1999) use random

spline curves to model individual player aging effects, though these too reflect a somewhat quadratic relationship between age and performance.

Including league as a fixed effect makes sense in that these are the only leagues we are interested in for this study. In this case the two MLB leagues American League (AL) and National League (NL), and two NPB leagues Central League (CL) and Pacific League (PL). We distinguish between these four because of the differing rules and styles of play; the AL and PL both have the Designated Hitter rule, while the NL and CL do not.

Finally, we would expect that position could explain some variability, especially given a typical batting line for a pitcher. For the analysis, since some of the player-seasons do not distinguish player's primary position outside of infield, outfield, designated hitter or pitcher, those are the four levels of the covariate we utilize, where: Infield (C, 1B, 2B, SS, 3B), Outfield (LF, CF, RF), Designated Hitter, and Pitcher. An excerpt of the data set, showing a single player's covariate and response values, is provided in Table 1.

3 Methodology

3.1 Model

There have been several studies examining streaky hitting in baseball (Albright 1993; Albert 2008, 2013) with the general consensus that there is not a great deal of difference between a player's behavior and a model that treats a player's seasonal plate appearances as a set of independent trials. Thus, we might assume we have vectors of count data for players $i = 1, \dots, m$ over seasons $j = 1, \dots, n_i$:

$$W_{ij} \sim \text{Multi}_{d+1}(PA_{ij}, \Pi_{ij})$$

Each vector W_{ij} is length $d + 1$, the number of outcome categories. We denote the exposure (i.e., plate appearances) for a player-season as PA_{ij} , so that we can express the total number of count vectors as $N = \sum_{i=1}^m n_i$. Our desire is to predict the counts of a new player i' for season $j = n_i + 1$, the first season after switching leagues. In our data, we have $m = 605$ players, $N = 4991$ player-seasons, and $d + 1 = 10$ categories.

Multinomial logistic regression, also known as the multinomial logit, is the most common form of analysis with a multinomial distributed response. The multinomial logit is an extension of binary logistic regression, and for a response vector of length $d + 1$, can in its most basic form be formulated as d "independent" logistic regressions with one category serving as a reference category.

Table 1: Data for Doug Ault who transitioned to Japan's Central League as an outfielder after four seasons as an infielder in the American League. Height, weight and age covariates have been standardized. Note some artifacts of different data sources; NPB data set had Ault as 194 cm tall instead of 191 cm, and 94 kg instead of 90.

Int	Height	Weight	Age	Age ²	Left	Switch	CL	NL	PL	OF	DH	P	1B	2B	3B	HR	SH	SF	BB	HBP	SO	OIP
1	1.05	0.25	-0.69	-0.71	0	0	0	0	0	0	0	0	5	1	0	0	0	0	1	0	3	11
1	1.05	0.25	-0.45	-0.49	0	0	0	0	0	0	0	0	73	22	3	11	3	3	39	4	68	268
1	1.05	0.25	-0.21	-0.27	0	0	0	0	0	0	0	0	20	1	1	3	2	0	17	1	14	65
1	1.05	0.25	0.27	0.20	0	0	0	0	0	0	0	0	19	5	1	3	0	1	14	2	23	93
1	1.59	0.70	0.51	0.44	0	0	1	0	0	1	0	0	70	12	0	18	2	2	14	4	60	166

The multinomial distribution, however, has the unfavorable property that there is a negative correlation between each pair of outcome counts. This is problematic because there are several outcomes in baseball that are known to be positively correlated. In addition, although we use covariates to allow the mean probability to differ across player-seasons, we expect there to be overdispersion. One model commonly used to account for overdispersion is the multinomial Dirichlet, but this model still imposes negative correlation. To account for overdispersion and allow for some positive correlations, we consider a compound distribution for the data, the multinomial logistic-normal distribution.

Specifically, we assume the following hierarchical structure for player i and season j :

$$W_{ij}|\Pi_{ij} \sim \text{Multi}_{d+1}(PA_{ij}, \Pi_{ij})$$

where

$$\Pi_{ij} \sim f(\Pi_{ij})$$

and $f(\Pi_{ij})$ is the logistic-normal distribution. Using Y_{ij} to represent the vector of log-odds:

$$Y_{ij} = \left(\log\left(\frac{\pi_{ij1}}{\pi_{ij(d+1)}}\right), \dots, \log\left(\frac{\pi_{ijd}}{\pi_{ij(d+1)}}\right) \right) \tag{1}$$

We incorporate the covariates through the linear model:

$$Y_{ij} = X_{ij}\beta + \psi_i + \epsilon_{ij} \tag{2}$$

with $\psi_i \sim N_d(0, \Phi)$, and $\epsilon_{ij} \sim N_d(0, \Sigma)$.

The random effect term, ψ_i , is a d -dimensional vector included to account for the longitudinal nature of the data (i.e., consecutive seasons from the same player). The inclusion of the error term ϵ_{ij} , and subsequently the covariance matrix Σ , differentiates this model from the multinomial logistic, thereby allowing positive correlations among categories. A more detailed description of the resulting covariance structure is provided in Appendix A.

We take a Bayesian inferential approach to model estimation. For prior distributions, we consider an improper flat prior on the fixed effect matrix and non-informative inverse-Wishart priors on the two covariance matrices. These prior distributions, explicitly defined in Appendix A, are as disperse as possible, allowing the data to drive the posterior distributions. We utilize a Metropolis-within-Gibbs sampler algorithm for fitting the model, due to the intractability of the multinomial logistic-normal conditional posterior likelihood.

As a baseline model, the multinomial logistic (M-Logit) model will be compared to our multinomial logistic-normal

(MLN) model. The M-Logit may be formulated as the MLN above, but with the absence of the overall error term, ε_{ij} . In the Bayesian framework of estimation, the difference in our models amounts to the MLN adding a sampling step for the covariance matrix Σ . Mixed effects M-Logit models have been addressed previously in both frequentist (Hartzel, Agresti, and Caffo 2001; Hedeker 2003) and Bayesian (Pettitt et al. 2006) settings. To our knowledge, the Bayesian mixed effects MLN model presented here has never been used.

3.2 Inference

For the MLN model described in Section 3.1, there are five sets of parameters within four iterated Metropolis-within-Gibbs steps that are to be estimated. The details of the algorithm are provided in Appendix A, while a summary of the parameters follows:

- (1) Y_{ij} : the length d latent player-season log-odds vectors.
- (2) β : the $p \times d$ matrix of fixed effects of the covariates.
- (3) ψ_i : the length d player random effect vector.
- (4) Φ : the $d \times d$ random effect covariance matrix.
- (5) Σ : the $d \times d$ error covariance matrix.

The Metropolis-within-Gibbs sampler results in long chains of samples of the parameters. After discarding a burn-in and thinning, these chains can be considered independent samples from the posterior distributions of the parameters, and are subsequently used in prediction (see Section 3.3).

The M-Logit model parameters, which are the same as the MLN parameters aside from Σ , are also estimated via Metropolis-within-Gibbs. The details of the M-Logit algorithm follow those of the MLN, in Appendix A.

3.3 Prediction

Suppose an NPB team is in need of a catcher and is considering Kevin Plawecki, former Purdue catcher and current catcher with the Boston Red Sox. Because he has only played in the major leagues (and not NPB) we do not have information regarding his player-specific ability (ψ_i) and therefore cannot make an accurate prediction of his performance if he were to transition to NPB. This section discusses how we approach this given that we have already generated samples from the joint posterior distribution of β , Σ , and Φ .

To accurately predict this player, we need an estimate of the ψ_i . We could add his output vectors to our data set and rerun the Metropolis-within-Gibbs sampler, but this is

going to take time. Because we have samples from the posterior distributions of y_{ij} , β , ψ_i , Σ , and Φ , and we do not expect the inclusion of one player to affect these, a faster alternative is to rerun the Metropolis-within-Gibbs sampler holding the β , Σ , and Φ quantities as fixed. In other words, only estimating the latent variables and random effect for player i .

By cycling through joint samples of β , Σ , and Φ , we obtain a corresponding set of posterior values of the new player's random effect, ψ_i , which we may use predict his performance. This is a useful approach when we want to quickly generate predictions for new players. A comparison of the predictions based on refitting the full data set and this approach was made and showed little difference. Thus, one might consider updating the model annually and using this approach to make predictions for players of interest during the year.

In general, all predictions follow the framework of first simulating draws from the posterior distributions of the parameters β , ψ_i , Σ , and Φ . Then, use those samples to generate a posterior predictive distribution of new player-season log-odds y_{ij} based on the parameter samples and covariate vector x_{ij} , and subsequently sample a posterior predictive distribution of counts W_{ij} . When interested in only few new players, however, the above method of fixing the β , Σ , and Φ allows for faster sampling of the operative quantity, the new player's random effect, ψ_i .

The main advantage of the Bayesian hierarchical model in this context is that we have generated samples from a posterior predictive distribution which will allow us to create prediction intervals for the outcome probabilities of a new player-season. These intervals allow us to not only compare the reward (the posterior mean) but also the risk (the uncertainty) for players moving from one league to another. These intervals will likely be too wide to be informative on an individual player basis. Several papers have discussed that simultaneous confidence intervals about multinomial counts tend to be quite wide (Quesenberry and Hurst 1964; Sison and Glaz 1995), and our hierarchical model accounts for further variability to allow for positive correlation. However such intervals are a quite useful tool for comparison of two or more players, at least in terms of relative floor and ceiling of player ability across categories of interest.

Based on pairs of posterior samples, $t = 1, \dots, T$, a prediction interval for a new player-season vector may be constructed via selecting the 95% most likely random effects for a new player under the posterior distribution of random effects ψ_i^t , calculating the expected log-odds for the new player-season $X_{i(n_i+1)}\beta^t + \psi_i^t$, and finally transforming those into probabilities. Multiplying by the

exposure represents an interval for the vector of expected counts. The same process may also be used to create prediction intervals for singular metrics, such as batting average (AVG) or on-base plus slugging percentage (OPS), which might be of interest for comparisons sake as a summary of player ability.

3.4 Assessing model fit

3.4.1 Model fit and comparison

In traditional regression, model fit diagnostics are often based on defining a form of residual and assessing if it behaves as it should under the proposed model. In a Bayesian context, this can take the form of posterior predictive checks. Formally defined by Rubin (1984) and championed by Gelman, Meng, and Stern (1996), the general framework of posterior predictive checking follows, given parameter set θ , data Y , and posterior predictive distribution $p(Y^l|Y)$ where Y^l are replicated data:

- (1) Simulate θ from the posterior distribution $p(\theta|Y)$
- (2) Simulate Y^l from the predictive distribution $p(Y^l|\theta, Y)$
- (3) Compare Y to the Y^l data sets

The third step occurs by defining a distance measure to form a residual for the data model. For category response data, which exist on the simplex, $\mathbb{S}^{d+1}$, many traditional distance measures are misleading (Aitchison 1986).

In our application, we are interested in the joint behavior of the count vectors relative to the MLN model. Following the general framework of posterior predictive assessment defined above, because each count vector has a different posterior predictive distribution, we are interested in whether the data jointly fit these distributions specified by our hierarchical structure. We propose comparing the squared Mahalanobis distances (Mahalanobis 1936) of the counts in log-odds space with what is expected under the model. We consider the Mahalanobis distance because it takes into account the covariance structure.

Following our framework and notation, for player i season j , we generate L new count vectors $W_{ij}^1, \dots, W_{ij}^L$, under the fitted model and transform them to the log-odds scale:

$$\tilde{Y}_{ij}^l = \text{alr}(W_{ij}^l), l = 1, 2, \dots, L \quad (3)$$

We then calculate the squared Mahalanobis distances of these model-generated log-odds, $\{\tilde{Y}_{ij}\}$, and observed log-odds, Y_{ij}^{obs} , using:

$$\widehat{MD}_{ij}^{2(l)} = \left((\tilde{Y}_{ij}^l - \bar{\tilde{Y}}_{ij})^T \hat{\Sigma}_{\tilde{Y}_{ij}}^{-1} (\tilde{Y}_{ij}^l - \bar{\tilde{Y}}_{ij}) \right) \quad (4)$$

and

$$MD_{ij}^{2(obs)} = \left((Y_{ij}^{obs} - \bar{Y}_{ij})^T \hat{\Sigma}_{Y_{ij}}^{-1} (Y_{ij}^{obs} - \bar{Y}_{ij}) \right) \quad (5)$$

where $\bar{\tilde{Y}}_{ij}$ is the sample mean of model-generated log-odds and $\hat{\Sigma}_{\tilde{Y}_{ij}}$ is their sample covariance.

Using the model-generated distance values, $\widehat{MD}_{ij}^{2(l)}$, we generate the empirical cumulative distribution function, $\mathcal{F}$, and then compute the percentile of the observed distance, $\mathcal{F}(MD_{ij}^{2(obs)})$, after correcting for the discreteness of the multinomial counts and the fact that the empirical distribution is based on L samples. If the data fit the model, we would then expect these percentiles to be distributed uniformly. This process can be seen as generating a form of Dunn-Smyth residual (Dunn and Smyth 1996). Dunn and Smyth propose back-transforming these percentiles to standard normal values to serve as residuals, r_{ij} . Normal quantile plots and residual plots can be used to assess fit.

An R package, DHARMA (Hartig 2019), has been published which extends this idea to creating residuals for results from fitted generalized linear mixed models in a univariate context. This multivariate extension, using the squared Mahalanobis distance, is straightforward to implement and provides a way to assess model fit under both the MLN and the M-Logit models.

Information criterion have played a large role historically in the comparison of model fit. Deviance information criterion (*DIC*) is, as described in Gelman et al. (2004), a somewhat Bayesian version of the Akaike information criterion (*AIC*). We will use *DIC* as a way to compare the fit of the MLN model with standard M-Logit model.

3.4.2 Handling Zero Counts

In practical applications, some observations will include zero counts for one or more categories. This is only an issue for two aspects of this analysis: obtaining initial estimates of the log-odds, Y_{ij} , and when calculating the Mahalanobis values. Zero counts do not pose an issue with parameter estimation.

The correction we use instead is based on one proposed by Smithson and Verkuilen (2006). It is possible to compress the data symmetrically around $\frac{1}{K}$, penalizing vectors with less information. In other words, for a general count vector W of length K , with exposure $PA = \sum_{k=1}^K W_k$ and in the presence of zero counts:

$$W_k^{comp} = W_k \left(1 - \frac{1}{PA} \right) + \frac{1}{K} \quad (6)$$

To illustrate, for a simple three category example, the count vector $W = (2, 0, 3)$ would become the compressed $W^{comp} = (1.9\bar{3}, 0.\bar{3}, 2.7\bar{3})$.

4 Results

We fit both the MLN and M-Logit models to our data set, separated into a training set of 500 players and a prediction set of 105 player-seasons. We did this for both a simple three category (home runs, walks, other) model and a full 10 category model.

4.1 Model fit

The three category model serves as a test case to illustrate the benefits of the MLN model over the M-Logit, the details

of which are provided in Appendix B. However, since our main interest is the full 10 outcome set, which provides a more complete picture of player ability and allows for computation of summary statistics, we report those results here.

We begin by comparing model fit of the 10 category training set with MLN and M-Logit models with *DIC* and then the Dunn–Smyth residual plots (Figures 1 and 2). We find that the model fit diagnostics are generally inconclusive on the most appropriate fit for the 10 category results. The *DIC* supports the M-Logit as the more appropriate model, with values of 141057 for M-Logit and 145797.5 for MLN. *DIC* favors M-Logit largely due to the sizable difference in number of effective parameters; the MLN model had over four times as many as M-Logit (10439.6–2225.3).

The standardized MD^2 , Dunn–Smyth type residual normal QQ-plots for the two model fits do not support one model over the other (Figures 1 and 2). Neither pointwise confidence envelope entirely includes the reference line that indicates adequate model fit. In some sense, it seems

Figure 1: Histogram and standard normal QQ-plot of standardized MD^2 percentile (Dunn–Smyth style) residuals for MLN fit, ten categories.

Figure 2: Histogram and standard normal QQ-plot of standardized MD^2 percentile (Dunn–Smyth style) residuals for M-Logit fit, ten categories.

the data falls somewhere between MLN and M-Logit; in other words, that some additional variability exists, but not as much as the MLN model is recovering. This suggests the bias we see in the MLN fit residuals is due to an overestimation of Σ , which will also be reflected in an excess of variability in predictions. In the future a more informative, less sparse, prior on Σ may be more reasonable here.

Since DIC favors the M-Logit while the residual plots seem to not favor either, we continue to consider both. In a statistical sense, it is generally prudent to be more conservative when estimating variability, as the MLN model seems to, especially when there is evidence that the alternative model, the M-Logit, underestimates it. However, our main goal is prediction, and we show in Section 4.3 that the MLN model does a slightly better job in that goal than the M-Logit.

4.2 Estimation

With $p = 13$ (including the intercept) and $d = 9$, the resulting β matrix contains 117 terms for each of the model fits. In the interest of brevity, we will highlight some of the more notable results from parameter estimation of the MLN fit.

The most interesting covariates for inference's sake are the aging curve and the league indicator variables. While the age effects are not among the largest in magnitude, they show some interesting trends, for instance that for a baseline player (average height, weight, right handed, American League, infielder) home run rate does not, on average, reach it's peak until past the age of 30 (Figure 3).

In terms of the league effect, we can summarize player ability via OPS and compare how a baseline player (average height, weight, age) is expected to perform in each league by position and handedness (Tables 2 and 3).

Figure 3: Posterior distribution (grey) and mean (black) of baseline player home run rate age trend per 500 plate appearances for MLN fit, ten categories.

Table 2: Right handed baseline player mean OPS based on recovered fixed effects for MLN fit of 10 category real training data set.

League	OPS (INF)	OPS (OF)	OPS (DH)	OPS (P)
American (MLB)	0.723	0.742	0.668	0.408
Central (NPB)	0.726	0.749	0.668	0.396
National (MLB)	0.722	0.741	0.669	0.417
Pacific (NPB)	0.778	0.802	0.720	0.428

Table 3: Left handed baseline player mean OPS based on recovered fixed effects for MLN fit of 10 category real training data set.

League	OPS (INF)	OPS (OF)	OPS (DH)	OPS (P)
American (MLB)	0.718	0.736	0.663	0.410
Central (NPB)	0.717	0.738	0.659	0.395
National (MLB)	0.717	0.735	0.665	0.418
Pacific (NPB)	0.768	0.790	0.710	0.427

Generally, we see that the NPB leagues (Central and Pacific) are easier to hit in than the MLB leagues (American and National), with the Pacific League producing the highest mean OPS across the board. What might be surprising is the relative weakness of the Designated Hitter position, and that the Pacific League pitchers (who do not usually bat, since the PL adopted the DH in analog with the AL) perform better on average than the Central League pitchers; as opposed to in the MLB where NL pitchers, who hit regularly, perform better than AL pitchers. It is also interesting that left-handed hitters in the CL do not appear to have quite the same advantage, when compared to the MLB leagues, as right-handed hitters.

Most of the covariance terms (26 of 36) recovered from Σ have posterior distributions which do not cover zero. Since the presence of Σ is the key distinction between MLN and M-Logit, allowing for positive correlation where M-Logit assumes negative correlation, investigating the implied correlation structure among the outcomes is telling of the MLN model's utility. Almost all of the recovered implied correlations between outcomes for a baseline player make practical sense. The strongest positive correlations are 0.19 between (HR, SF), (1B, OIP) and (HR, BB). The strongest within season correlation is the -0.69 between (SO, OIP), followed by the -0.52 between (1B, SO). Other correlations of note are the positive correlation between (HR, SO), 0.12, and the negative correlation between (3B, HR), -0.04 .

These make quite a bit of sense; striking out and putting the ball in play are exactly opposite outcomes, and

players who hit a lot of singles are likely putting the ball in play often as well, instead of striking out. Meanwhile, players often attempt to hit home runs when runners are on third base, assuming that even if they don't get all of the ball, a sacrifice fly will net the team a run. Players who hit home runs have also long been considered to also be walk and strikeout prone, while also tend to be slower and less likely to hit many triples.

Before moving to prediction of out-of-sample players, we note that there is some evidence of strong shrinkage of the recovery of random effects for those players with little information being used to estimate random effects. Comparing the estimated OPS to observed OPS for qualified batters in the initial model fit, there is a general underestimation (Figure 4). This is largely due to many players who switch from MLB to NPB with little information being predicted close to the mean due to strong shrinkage in random effect recovery, and then outperforming that prediction. However, comparing out-of-sample predictions of NPB to MLB players with an established prediction method, the Marcells, shows that even with this shrinkage, our method performs competitively (see Section 4.3.1).

Figure 4: OPS estimation for all qualified batters ($PA > 500$) based on MLN fit of training set, ten categories.

4.3 Prediction

To assess prediction, we focus on the players whose careers were held out of the model fit. Figure 5 contains three plots of observed versus predicted outcomes (on the square root scale) that vary in terms of frequency. More frequent outcomes—that is, 1B, 2B, HR, BB, SO and OIP—show a fairly tight spread around the 45° line. Less frequent outcomes—that is, 3B, SH, SF, HBP—show a tendency to underpredict. This is perhaps more clear in Figure 6 where we convert each observed count into a Dunn–Smyth type residual and plot them versus the fitted values. Overall, the model performed well with some slight issues predicting the rare outcomes.

The MLN also predicts slightly better than the M-Logit model. We attribute this to the added flexibility of the MLN to handle positively correlated outcomes. For example, the correlation between predicted and observed OPS for the held out players was 0.526 for MLN and 0.513 for the M-Logit.

As far as players transitioning between the leagues, Figure 7 shows the prediction intervals of OPS for four players along with their observed value. These players were randomly selected for the prediction set, but are chosen here for comparisons sake. Mike Young and Rod Allen both moved from the AL to CL at the age of 30, predominantly playing OF. However, Young played nine years in MLB with 235 PA/year as opposed to Allen, who played only three years with 18 PA/year. Thus they represent very similar players in terms of covariates, but disparate career lengths and information. Tsuyoshi Shinjo and Kosuke Fukudome both played many years in Japan before moving to MLB, with similar numbers, which on the surface may make choosing between them, should a team have to, more difficult.

Allen's prediction interval is wider than Young's, reflecting the additional uncertainty in recovery of his random effect. While Allen's OPS was not predicted well by the model fit, it was still within the rather wide prediction

Figure 5: Square root of predicted versus observed counts for three categories (3B, HR, SO) of test set players based on MLN fit, ten categories.

Figure 6: Standardized MD^2 percentile (Dunn–Smyth style) residual plots for three categories (3B, HR, SO) of test set players based on MLN fit, ten categories.

Figure 7: OPS prediction intervals with posterior means (black circle) and observed value (red cross) for four players under MLN fit, ten categories.

interval. Though Shinjo and Fukudome had similar careers in Japan, and a similar uncertainty about their predicted OPS, Fukudome had a higher floor, ceiling, and predicted OPS than Shinjo. Subsequently, he also performed at a higher level (according to OPS) than Shinjo in his first year in MLB.

There is quite a bit of uncertainty in forecasting a player's seasonal performances, somewhat limiting the usefulness of these intervals. However, for our purposes, these intervals can be used to provide a comparable level of risk associated with each player.

4.3.1 Comparison to the Marcells

Pre-2015, the time frame which our data set covers, there were only 50 players to move from NPB to MLB, 23 of which were position players. For players with no previous major league experience, such as NPB players moving to MLB, the Marcells predicts those players at a weighted league average of the three previous seasons, with an adjustment for age. While the Marcells will provide slightly different predictions for each player (given different age and season) there is no prior player performance information being used in generating the predictions. Since there have been no published NPB predictions, and for this project we did not have easy access to the full NPB data set so we cannot create a “Japanese version” of the Marcells, and comparing predictions for the 555 players to make the transition from MLB to NPB becomes infeasible.

We generate posterior predictive samples under the MLN model for the 23 position players to make the NPB to MLB transition, and calculate the posterior predictive mean of those samples as the point prediction to compare to the Marcells predictions for those same players in the first season after making the transition. One player (Andy Abad for the 2001 Athletics) had only a single plate appearance in their first year in MLB, so we ignore him and focus on the other 22. Figures 8–10 plot the observed versus predicted counts (on the square root scale) for both the MLN model

Figure 8: Comparison of MLN (left panel) and Marcells (right panel) predicted triples and for 23 position players to make transition from NPB to MLB.

Figure 9: Comparison of MLN (left panel) and Marcells (right panel) predicted home runs and for 23 position players to make transition from NPB to MLB.

Figure 10: Comparison of MLN (left panel) and Marcells (right panel) predicted strikeouts and for 23 position players to make transition from NPB to MLB.

and Marcells in three of the outcomes. In general, the MLN model provided a tighter fit around the 45° line. Table 4 summarizes the correlation and predicted residual error sum of squares (PRESS) statistic for each of the 10 outcomes plus four other common statistics.

Nearly across the board, the model outperforms the Marcells in terms of prediction. Most count outcomes, and all summary statistics were better predicted under the MLN model than by the Marcells. Considering the Marcells contains no prior player performance in its calculation, this is perhaps a low bar to clear, but it is an established bar. Indeed, the Marcells has generally been shown to be competitive with more complicated prediction models (Larson 2015). While first year predictions of past Japanese players under the other popular projection systems are not available, by outperforming the Marcells here it suggests that MLN may be competitive with, or outperform, these other established models.

5 Conclusions

Results show that the methodology is promising in its ability to provide accurate predictions for baseball player seasons after switching leagues, while also indicate there is still work to be done in refining the model. There is evidence that the standard multinomial logistic model does

not adequately account for the overdispersion or, critically, positive correlations that occur in baseball data. When comparing models in terms of DIC, we found that the dimensionality of the response had an impact on whether the multinomial logistic-normal model was preferred to the simpler M-Logit. The Dunn–Smyth residual QQ-plots also varied by granularity, and did not provide a clear delineation of model fit for the 10 category model. In terms of prediction, the model performs quite well and manages to outperform the Marcells—one of the most commonly used projection systems. For this key reason, and the added flexibility gained from the inclusion of the Σ covariance matrix, the model is useful in this context. Additionally, the framework developed for providing uncertainty quantification for predictions of baseball seasonal count outcomes is a novel contribution to the study of baseball performance.

The estimation results coincide with traditional thinking about the difference in leagues; that it is easier to succeed as a power hitter in the Japanese Leagues. In part because of this, there could be a bias introduced in terms of a player's estimated random effect when only estimated based on their performance in another league. This is a limitation in using a single random effect vector to describe a player's performance relative to the mean performance. A potential option would be to include a fixed inflation factor attached to league that is modeled as part of a player's random effect after they switch leagues. This would

Table 4: Pairwise correlations and PRESS statistics for MLN model predicted values and observed values, and between the Marcells and observed values, for 10 categories of interest and summary statistics batting average, on-base percentage, slugging percentage, and on-base plus slugging percentage.

CORR														
	1B	2B	3B	HR	SH	SF	BB	HBP	SO	OIP	AVG	OBP	SLG	OPS
MLN	0.991	0.972	0.708	0.941	0.788	0.891	0.956	0.875	0.970	0.995	0.674	0.291	0.449	0.368
Marcls	0.968	0.959	0.714	0.893	0.675	0.885	0.882	0.625	0.886	0.965	0.379	0.160	0.278	0.270
PRESS														
	1B	2B	3B	HR	SH	SF	BB	HBP	SO	OIP	AVG	OBP	SLG	OPS
MLN	85.670	23.420	3.956	3.965	2.983	1.088	56.351	3.246	96.235	209.540	0.005	0.012	0.023	0.053
Marcls	289.109	17.669	4.401	23.151	4.785	1.137	138.889	6.768	409.945	3521.136	0.008	0.014	0.027	0.063

account for the direction of a player's move and serve to mitigate some of the presumed bias we see in the recovery of the random effects. Another potential correction, which would be one way to allow the model to handle inclusion of players who did not switch leagues, would be to include two new indicator variables which correspond to a player being in a "foreign" league. This would help mitigate the potential further bias in a player's random effect that would likely occur with the inclusion of players who do not switch leagues.

Author contribution: All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.

Research funding: None declared.

Conflict of interest statement: The authors declare no conflicts of interest regarding this article.

Appendix A Multinomial logistic-normal model

A.1 Model details

While the full model is provided in Section 3.1, we provide some additional background, properties, and discussion of the model in this appendix. To determine a reasonable distribution $f(\Pi_{ij})$, we turn to compositional data analysis. The modern discussion of modeling compositional data begins with John Aitchison, who in the 1980s developed many of the tools used today. In his original works Aitchison (1982, 1986) discusses two main classes of models for modeling compositional data, the Dirichlet and logistic-normal classes. Aitchison leans towards the logistic-normal approach (Aitchison 1982) due to the strong internal independence structure of the Dirichlet. In other words, the multinomial-Dirichlet model has the same negative correlation structure as the multinomial.

The covariance matrix, Σ , arises from modeling the probabilities with the logistic-normal. It represents the covariance structure of the underlying log-odds. It is used to allow for positive correlations among outcome categories; our solution to the same problem Null (2009), Albert (2016), and Powers, Hastie, and Tibshirani (2018) identified, and used nesting structures of outcomes to solve. In a mixed model framework, we also include a single vector of random effects for each player, which produces another covariance matrix, Φ , which accounts for the longitudinal nature of the data.

It is straightforward to allow for the inclusion of multiple player-specific random effect vectors, but for our simplicity we focus on a single random intercept vector.

Because the log-odds Y_{ij} can be considered multivariate normal latent variables, the inclusion of the random intercept allows us to account for the correlation in each outcome across a player's career, while the error term, ε_{ij} , builds in correlation between outcomes within player-seasons. This is better visualized by the distribution of all sets of latent log-odds vectors from player i :

$$Y_i|X_i, \theta \sim N_{dn_i} \left(\begin{bmatrix} \vec{x_{i1}}\beta \\ \vdots \\ \vec{x_{in_i}}\beta \end{bmatrix}, \begin{bmatrix} \Phi + \Sigma & \dots & \Phi \\ \vdots & \ddots & \vdots \\ \Phi & \dots & \Phi + \Sigma \end{bmatrix} \right) \quad (7)$$

where $\theta = (\beta, \Sigma, \Phi)$.

This covariance structure is significantly more flexible than that of the traditional mixed multinomial logistic model, which does not directly model the correlation in outcomes within player-seasons.

A.2 Bayesian prior selection and Gibbs sampler algorithm

We have the known quantities W_{ij} , PA_{ij} and X_{ij} , and model parameters Π_{ij} , Y_{ij} , β , ψ_i , Σ , and Φ . For prior distributions, we consider an improper flat prior on the fixed effect matrix and noninformative inverse-Wishart priors on the two covariance matrices:

$$\begin{aligned} p(\beta) &\propto 1 \\ \Sigma &\sim \text{Wishart}^{-1}(v_1 = d, \Lambda_1 = I_d + 1_d 1_d^T) \\ \Phi &\sim \text{Wishart}^{-1}(v_2 = d, \Lambda_2 = I_d + 1_d 1_d^T) \end{aligned}$$

Though the issue of separation can be problematic with the choice of a flat prior for β , this was not an issue with this data set nor any of the simulated data sets. If separation is an issue, a proper prior for the fixed effect matrix, such as the matrix normal distribution, should be used.

We utilize a Metropolis-within-Gibbs sampler algorithm, where samples are drawn iteratively from conditional posterior distributions of Y_{ij} , β , ψ_i , Σ , and Φ . We first describe the framework for the Gibbs sampler, allowing for a single random intercept ($q = 1$). When multiple random effects are included in the model ($q > 1$), some of the expressions are slightly more complicated. These expressions are presented after the simpler single random intercept framework.

- (1) Initialize starting values, $t - 1 = 0$, Y_{ij}^0 , β^0 , ψ_i^0 , Φ^0 and Σ^0 .
- (2) Update of Y_{ij} (Metropolis step):

$$\begin{aligned} Y_{ij}^t | X_{ij}, \beta^{t-1}, \psi_i^{t-1}, \Sigma^{t-1}, W_{ij} &\propto p(W_{ij} | Y_{ij}^t) \\ &\times f(Y_{ij}^t | X_{ij}, \theta_i^{t-1} = (\beta^{t-1}, \psi_i^{t-1}, \Sigma^{t-1})) \\ p(W_{ij} | Y_{ij}^t) &\times f(Y_{ij}^t | X_{ij}, \theta_i^{t-1}) \sim \text{Multi}_{d+1}(PA_{ij}, \prod(Y_{ij}^t)) \\ &\times N_d(X_{ij}\beta^{t-1} + \psi_i^{t-1}, \Sigma^{t-1}) \end{aligned}$$

where Y_{ij}^t is proposed from a normal approximation to the beta distribution.

- (3) Joint Update of β , ψ_i (Gibbs step):

$$\begin{aligned} \beta^t, \psi_i^t | X_i, \Sigma^{t-1}, \Phi^{t-1}, Y_i^t &\propto f(\beta^t | X_i, \Sigma^{t-1}, \Phi^{t-1}, Y_i^t) \\ &\times f(\psi_i^t | \beta^t, X_i, \Sigma^{t-1}, \Phi^{t-1}, Y_i^t) \end{aligned}$$

- (a) Update of β without conditioning on ψ_i :

$$\text{vec}(\beta^t) | X_i, \Sigma^{t-1}, \Phi^{t-1}, Y_i^t \sim N_{d(p+1)}(\mu_\beta = \text{vec}(A^{-1}b), \Sigma_\beta = A^{-1})$$

where:

$$A = \sum_{i=1}^m \left((X_i^T X_i) \otimes (n_i \Phi^{t-1} + \Sigma^{t-1})^{-1} \right) \quad (8)$$

$$b = A \left(\sum_{i=1}^m X_i^T X_i \right)^{-1} \left(\sum_{i=1}^m X_i^T Y_i^t \right) \quad (9)$$

- (b) Update of ψ_i :

$$\text{vec}(\psi_i^t) | \beta^t, X_i, \Sigma^{t-1}, \Phi^{t-1}, Y_i^t \sim N_d(\tilde{\psi}_i, U_i)$$

where:

$$\tilde{\psi}_i = U_i (\Sigma^{(t-1)^{-1}} \otimes 1_{n_i}) \text{vec}(Y_i^t - X_i \beta^t) \quad (10)$$

$$U_i = (\Phi^{(t-1)^{-1}} + (\Sigma^{(t-1)^{-1}} \otimes 1_{n_i} 1_{n_i}^T))^{-1} \quad (11)$$

- (4) Update of Φ (Gibbs step):

$$\Phi^t | \Psi^t \sim \text{Wishart}^{-1}(v_2' = v_2 + m, \Lambda_2' = \Lambda_2^{-1} + \Psi^t \Psi^{tT})$$

where:

$$\Psi^t = (\psi_1^t, \dots, \psi_m^t)^T \quad (12)$$

- (5) Update of Σ (Gibbs step):

$$\Sigma^t | \Psi^t, Y^t \sim \text{Wishart}^{-1}(v_1' = v_1 + N, \Lambda_1' = \Lambda_1^{-1} + \tilde{e} \tilde{e}^T)$$

where:

$$Y^t = (Y_1^t, \dots, Y_m^t)^T \quad (13)$$

$$\tilde{e}_i = Y_i^t - X_i \hat{\beta}^t - \psi_i^t \quad (14)$$

$$\hat{\beta}^t = (Y^t - Z \Psi^t) X^T (X X^T)^{-1} \quad (15)$$

- (6) Set $t = t + 1$, repeat steps (2)–(5) until convergence criteria met.

When multiple random effects are of interest, it is straightforward to define a random covariate design matrix, Z , made up of vectors Z_i for players $i = 1, \dots, m$. Most of the above algorithm can be represented the same way with this simple inclusion. However, one of the steps is complicated. The expression in step (3)(a), the posterior of β without conditioning on ψ_i , is a little more complex. Equation (8) must be rewritten as an expression where 1_{ni} is replaced by Z_i :

$$A = \sum_{i=1}^m (X_i \otimes I_d)^T Q_i (X_i \otimes I_d) \quad (16)$$

$$Q_i^{-1} = (Z_i \otimes I_d) \Phi^{(t-1)} (Z_i \otimes I_d)^T + (I_{n_i} \otimes \Sigma^{(t-1)}) \quad (17)$$

A.2.1 M-Logit Gibbs sampler

We describe a Metropolis-within-Gibbs sampler for fitting the mixed-effects M-Logit model with a single random effect. This algorithm is used in comparing model fit between data simulated under the two data archetypes, MLN versus M-Logit.

- (1) Initialize starting values, $t - 1 = 0$, μ^0 , β^0 , ψ_i^0 , and Φ^0 .
 (2) Update of ψ_i :

$$\begin{aligned} \psi_i^t | X_i, \mu^{t-1}, \beta^{t-1}, \Phi^{t-1}, W_i &\propto p(W_i | X_i, \mu^{t-1}, \beta^{t-1}, \psi_i^t) \\ &\times f(\psi_i^t | X_{ij}, \theta_i^{t-1} = (\mu^{t-1}, \beta^{t-1}, \Phi^{t-1})) \end{aligned}$$

via Metropolis algorithm.

$$\begin{aligned} p(W_i | X_i, \mu^{t-1}, \beta^{t-1}, \psi_i^t) &\times f(\psi_i^t | X_{ij}, \theta_i^{t-1}) \\ &\sim \prod_{j=1}^{n_i} n_i \text{Multi}_{d+1}(PA_{ij}, \prod(Y_{ij}^t)) \times N_d(\mu^{t-1}, \Phi^{t-1}) \end{aligned}$$

Where $Y_{ij}^t = \mu^{t-1} + X_{ij}\beta^{t-1} + \psi_i^t$, and

$$\prod(Y_{ij}^t) = \left(\frac{e^{y_{ij1}^t}}{1 + \sum_{k=1}^d e^{y_{ijk}^t}}, \dots, \frac{1}{1 + \sum_{k=1}^d e^{y_{ijk}^t}} \right) \quad (18)$$

With multivariate normal proposal distribution:

$$\psi_i^t \sim N_d(\psi_i^{t-1}, c_\psi \Phi^{t-1})$$

Where c_ψ is a scalar used to control the “step-size” of the proposal distribution.

- (3) Update of Φ :

$$\Phi^t | \Psi^t \sim \text{Wishart}^{-1}(\nu_2' = \nu_2 + m, \Lambda_2' = \Lambda_2^{-1} + \Psi^t \Psi^{tT})$$

Where:

$$\Psi^t = (\psi_1^t, \dots, \psi_m^t)^T \quad (19)$$

- (4) Update of μ :

$$\mu^t | \Psi^t, \Phi^t \sim N_d\left(\bar{\Psi}^t, \frac{\Phi^t}{m}\right)$$

Where:

$$\bar{\Psi}^t = \frac{1}{m} \sum_{i=1}^m \psi_i^t \quad (20)$$

- (5) Update of β :

For each vector β_k in the matrix β update given, β_{-k} the matrix with column k removed, corresponding to the conditional likelihood (Holmes and Held 2006):

$$l(\beta_k^t | \beta_{-k}^{t-1}, W_{ijk}, \psi_i^t) = \prod_{i=1}^m \prod_{j=1}^{n_i} \left(\frac{e^{\eta_{ijk}^{t-1}}}{1 + e^{\eta_{ijk}^{t-1}}} \right)^{W_{ijk}} \left(\frac{1}{1 + e^{\eta_{ijk}^{t-1}}} \right)^{PA_{ij} - W_{ijk}} \quad (21)$$

Where:

$$\eta_{ijk}^{t-1} = X_{ij}\beta_{-k}^{t-1} + \psi_i^t - C_{ijk}^{t-1} \quad (22)$$

$$C_{ijk}^{t-1} = \log\left(1 + \sum_{l \neq k} e^{X_{ij}\beta_l^{t-1} + \psi_i^t}\right) \quad (23)$$

via Metropolis algorithm with multivariate normal proposal distribution of dimension p , the number of covariates (not including the intercept):

$$\beta_k^t \sim N_p(\beta_k^{t-1}, c_\beta \Sigma_\beta^{t-1})$$

Where c_β is a scalar used to control the “step-size” of the proposal distribution Σ_β^{t-1} can be an proposal covariance matrix. One option is to use Pólya-Gamma auxiliary variables Polson, Scott, and Windle (2013) to estimate a reasonable shape of the covariance matrix:

$$\sum_{\beta}^{t-1} = X^T \Omega_k^{t-1} X \quad (24)$$

Where $\Omega_k^{t-1} = \text{diag}(\{\omega_{ijk}^{t-1}\})$, and ω_{ijk}^{t-1} are Pólya-Gamma auxiliary variables given β_k^{t-1} :

$$\omega_{ijk}^{t-1} \sim PG(PA_{ij}, \eta_{ijk}^{t-1})$$

- (6) Set $t = t + 1$, repeat steps (2)–(5) until convergence criteria met.

Appendix B Three category results

While many different levels of granularity may be defined for baseball data, we focused on 10 categories for our main results, since those were the outcomes necessary for construction of summary statistics such as OPS. One of the drawbacks of this, however, was that the benefits of our model were not as clear. The key usefulness of our MLN model over the traditional M-Logit model is the additional flexibility in accounting for positive correlations between outcomes. With 10 categories, this implies 45 pairwise correlations, many of which will be negative. We found that our MLN model still produces slightly better predictions than the M-Logit in the 10 category model. The model fit diagnostics, conversely, are inconclusive on the benefit of allowing for this additional flexibility in modeling covariance when so many of the covariance terms are negative. We believe that since the M-Logit model still shows evidence of underestimating variability, and our model predicts slightly better than the M-Logit, that our model is still preferred even in the 10 category case.

However, a granularity of three outcomes, where we choose two of the three categories to be positively correlated, serves as a slightly stronger test case for the benefits of the MLN model. For the results reported our three categories will be Home Runs, Base on Balls, and everything else; (*HR*, *BB*, *Other*). Home runs and base on balls (also known as walks) are known to be positively correlated. Due to the unit sum constraint, the positive correlation between home runs and walks necessarily implies negative correlations between home runs and other, and walks and other. We once again separate our data into a training set of 500 players and a test set of 105 players, using all of the same covariates as described in Section 2.2. We fit the training set with both the MLN and M-Logit models, and the use the test set to generate predictions under each fit.

We proceed by comparing and discussing the model fit, estimation, and prediction of both our MLN model and the M-Logit model for this three category case.

B.1 Model fit

Our two main methods of comparing model fit are the Deviance Information Criterion and the Dunn–Smyth residual plots. In this case, both methods show a preference for the MLN model. In terms of DIC (Table 5) the ratio in number of effective parameters stays about the same as in the 10 category model, with the MLN model

involving almost four and a half times the effective number of parameters. However, in the three category model, this disparity is not enough to washout the benefits in maximizing the log-likelihood. With the lower DIC, this criterion supports our MLN model as more appropriate for the data.

The Dunn–Smyth type residual normal QQ-plots likewise show a distinctly better fit of the data to the MLN model (Figures 11 and 12). As the estimation and prediction results will also show, the M-Logit model struggles to account for the positive correlation between home runs and walks, and does not adequately model the additional variability that is present in the data.

B.2 Estimation

The three category application produces a covariate effect matrix β with dimension 13×2 (including the intercept), making direct comparison of the recovery of these matrices between the MLN and M-Logit models more straightforward than in the 10 category models. While we cannot know the true values of these fixed effects, there are some notable differences in the recovery of β . Since the only difference in the two models is the additional variance term in the MLN model, this difference provides some evidence that the M-Logit is perhaps producing bias in the recovery of the fixed effects in order to account for the positive correlation in the data. Consider the 12 covariate effect vectors for the MLN (Table 6) and M-Logit (Table 7) models.

Each vector of length $d = 2$ corresponds to the change in log-odds for a one unit increase in the corresponding covariate above the baseline. Vectors with elements near zero indicate the covariate does not have a large effect; in both models the largest effect is from the indicator variable for pitchers, which makes a great deal of sense. A pitcher hitting will be much less likely to both hit a home run and walk; hence the large negative terms. This effect is captured exactly the same (to two decimal places) by both models. However, some of the other effects show large differences. For instance, the three league indicator variables, which represent switching from the American

Table 5: Effective number of parameters and deviance information criterion for three category real data training set fit under MLN and M-Logit models.

	MLN Model Fit	M-Logit Model Fit
p_{DIC}	3077.88	687.96
DIC	37442.81	38387.73

Figure 11: Histogram and standard normal QQ-plot of standardized MD^2 percentile (Dunn–Smyth style) residuals for MLN fit, three categories.

Figure 12: Histogram and standard normal QQ-plot of standardized MD^2 percentile (Dunn–Smyth style) residuals for M-Logit fit, three categories.

Table 6: Recovery of fixed effect (β) matrix under MLN fit of 3 category real training data set.

	Height	Weight	Age	Age ²	Left	Switch	CL	NL	PL	OF	DH	P
$\log\left(\frac{HR}{Other}\right)$	0.14	0.16	0.29	-0.22	-0.05	-0.08	0.42	-0.13	0.42	0.14	-0.12	-2.00
$\log\left(\frac{BB}{Other}\right)$	0.03	0.05	0.31	-0.21	0.11	0.22	-0.13	-0.02	-0.01	-0.07	-0.10	-1.00

Table 7: Recovery of fixed effect (β) matrix under M-Logit fit of 3 category real training data set.

	Height	Weight	Age	Age ²	Left	Switch	CL	NL	PL	OF	DH	P
$\log\left(\frac{HR}{Other}\right)$	0.14	0.04	0.62	-0.62	-0.05	-0.20	0.62	0.02	0.61	0.09	-0.12	-2.00
$\log\left(\frac{BB}{Other}\right)$	0.04	0.00	0.28	-0.25	0.10	0.17	0.06	0.10	0.15	0.00	0.10	-1.00

League to one of the Central, National, or Pacific Leagues, all differ in the effect of switching on walks. In the MLN model, the negative signs for the elements corresponding to $\log\left(\frac{BB}{Other}\right)$ for the three league covariates indicate a decrease in walk rate when switching from the AL, while the signs for all three terms in the M-Logit model are positive.

Since we do not know the true value of these terms, it is impossible to truly know which model is closest to the truth. However, as the prediction results show, the MLN model seems to do as well of a job of predicting the home run and other categories, and to a better job predicting the walks category, than M-Logit.

B.3 Prediction

Comparing the correlation between predicted and observed outcomes for both model fits shows the MLN perform barely better in prediction for all three categories. While the difference in correlation for home runs and other is negligible, there is a slightly more sizeable difference in the prediction of walks (Table 8). Similarly to the 10 category results, the improvement in prediction of the MLN model over the M-Logit model seems to be marginal at best on the individual category level, especially considering we only had time for a single fold cross-validation. At the very least, the MLN model seems to do no worse than the M-Logit model in terms of prediction. Coupled with the clear benefit in terms of model fit at this granularity, there is evidence in this application that the MLN model is the more appropriate model.

Appendix C The Marcells

Due to the proprietary nature of many of the most popular prediction methods currently in use, there are limited options for comparing the performance of our model; especially for those players who transitioned many years ago. Further, to our knowledge there has to this point been no effort to predict player performance in the NPB, so we are restricted to only those making the transition to MLB.

Table 8: Correlation between predicted and observed outcome counts of test set under both MLN and M-Logit model fits.

	HR	BB	Other
MLN	0.8389	0.8818	0.9974
M-Logit	0.8386	0.8682	0.9971

Using the full MLB data set available in the Lahman package (Friendly 2015) in R, it is possible to produce Marcells predictions for a position player of any age in any year in MLB. The Marcells consist of a relatively simple ratio of a weighted mean of a player's performance for the past three years and the league average over that time, with a piece-wise adjustment for age.

The weighting scheme the Marcells use is 3–4–5 for the previous three seasons, motivating the 1200 plate appearances as if adding 300, 400, and 500 league average plate appearances to the player's own weighted plate appearances to form the prediction. In terms of uncertainty, the Marcells provides a singular value which describes the “reliability” of a prediction; simply the proportion of the prediction which comes from player data as opposed to the regression to the mean. For example, a player who had a weighted total of 7000 plate appearances over three years would have a reliability score of $\frac{7000}{7000+1200} = 0.854$. Players with no data, including minor league players and players coming from Japan such as in our data set, will be predicted at the mean and have a reliability score of 0 (Tango 2012).

To compute the Marcells prediction for a player i in season j , the major league average probability vectors $\bar{\Pi}$ are calculated from all position players for seasons $j-3$, $j-2$, and $j-1$. The unadjusted probability vector for season j is then:

$$\prod_j^{(m)} = \frac{3\bar{\Pi}_{j-3} + 4\bar{\Pi}_{j-2} + 5\bar{\Pi}_{j-1}}{12} \quad (25)$$

If a player has played in any of the previous three seasons, $\Pi_j^{(m)}$ is multiplied by 1200 and then added to the player's weighted counts (call W_{ij} the count vector of player i in season j) over the last three years, then the player's unadjusted probability vector is computed by dividing the weighted counts by the sum of the weighted observed plate appearances and 1200:

$$\prod_{ij}^{(m)} = \frac{3W_{ij-3} + 4W_{ij-2} + 5W_{ij-1} + 1200(\Pi_j^{(m)})}{3PA_{ij-3} + 4PA_{ij-2} + 5PA_{ij-1} + 1200} \quad (26)$$

The Marcells then employs a simple age adjustment for ages over 29 years by a piecewise linear function:

$$\prod_{ij}^{(ma)} = \begin{cases} \prod_{ij}^{(m)} (1 + 0.006(29 - a_{ij})) & a_{ij} \leq 29 \\ \prod_{ij}^{(m)} (1 + 0.003(29 - a_{ij})) & a_{ij} > 29 \end{cases}$$

To obtain the predicted count vector, $\hat{W}_{ij}^{(ma)}$, the probability vector is multiplied by the predicted number of plate appearances. For players with no prior information, the

Marcells uses $PA = 200$, but in our case we know the observed number of plate appearances for the player in the season being predicted, PA_{ij} .

$$\hat{W}_{ij}^{(ma)} = PA_{ij} \Pi_{ij}^{(ma)} \quad (27)$$

The Marcells reliability score, $r_{ij}^{(m)}$, is computed simply by the ratio of weighted plate appearances from the player's previous three years, to the the sum of those weighted plate appearances and the 1200 league average plate appearances:

$$r_{ij}^{(m)} = \frac{3PA_{ij-3} + 4PA_{ij-2} + 5PA_{ij-1}}{3PA_{ij-3} + 4PA_{ij-2} + 5PA_{ij-1} + 1200} \quad (28)$$

illustrating how players without any prior data, such as minor league players or Japanese players, have a reliability score of 0. Since a player rarely achieves more than 750 plate appearances in a season, the maximum reliability score for a Marcells prediction is approximately 0.88. This score can be interpreted as a measure of "confidence" in a prediction, though it is comparable across players only insofar as being related to the relative number of plate appearances two players have had over the past three years.

Acknowledgements: Aggregated NPB data was provided by Data Stadium, Inc., S-GATE Akasaka, 6-2-4, Akasaka, Minato-ku, Tokyo 107-0052, while MLB data was obtained via the `\pkg{Lahman}` package in `\proglang{R}` `\citep{Lahman:2015r}`. We would like to thank Dr. Jim Albert for connecting us with Data Stadium in the first place and providing inspiration for this project. All data is otherwise available publicly online, on disparate pages, from baseball-reference.com.

References

- Aitchison, J. 1982. "The Statistical Analysis of Compositional Data." *Journal of the Royal Statistical Society: Series B* 44 (2): 139–77.
- Aitchison, J. 1986. *The Statistical Analysis of Compositional Data*. 29 West 35th Street. New York, NY 10001: Chapman & Hall.
- Albert, J. 1994. "Exploring Baseball Hitting Data: What about Those breakdown Statistics?" *Journal of the American Statistical Association* 89 (427): 1066–74.
- Albert, J. 2002. *Smoothing Career Trajectories of Baseball Hitters*. Technical report. Bowling Green, Ohio, USA: Department of Mathematics and Statistics, Bowling Green State University.
- Albert, J. 2008. "Streaky Hitting in Baseball." *Journal of Quantitative Analysis in Sports* 4 (1). Article 3, <https://doi.org/10.2202/1559-0410.1085>.
- Albert, J. 2013. "Looking at Spacings to Assess Streakiness." *Journal of Quantitative Analysis in Sports* 9 (2): 151–63.
- Albert, J. 2016. "Improved Component Predictions of Batting and Pitching Measures." *Journal of Quantitative Analysis in Sports* 12 (2): 73–85.
- Albright, S. C. 1993. "A Statistical Analysis of Hitting Streaks in Baseball." *Journal of the American Statistical Association* 88 (424): 1175–83.
- Bennett, J. M., and J. A. Flueck. 1983. "An Evaluation of Major League Baseball Offensive Performance Models." *The American Statistician* 37 (1): 76–82.
- Berry, S. M., C. S. Reese, and P. D. Larkey. 1999. "Bridging Different Eras in Sports." *Journal of the American Statistical Association* 194 (447): 661–76.
- Chandler, G., and G. Stevens. 2012. "An Exploratory Study of Minor League Baseball Statistics." *Journal of Quantitative Analysis in Sports* 8 (4): 1–26.
- Dunn, K. P., and G. K. Smyth. 1996. "Randomized Quantile Residuals." *Journal of Computational and Graphical Statistics* 5: 1–10.
- Efron, B., and C. Morris. 1975. "Data Analysis Using Stein's Estimator and its Generalizations." *Journal of the American Statistical Association* 70: 311–9.
- Fair, R. C. 2008. "Estimating Age Effects in Baseball." *Journal of Quantitative Analysis in Sports* 4 (1). Article 1, <https://doi.org/10.2202/1559-0410.1074>.
- Friendly, M. 2015. *Lahman: Sean Lahman's Baseball Database*. R package version 4.0-1.
- Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. 2004. *Bayesian Data Analysis. The Edinburgh Building*. Cambridge CB2 8RU, UK: Chapman & Hall/CRC.
- Gelman, A., X.-L. Meng, and H. Stern. 1996. "Posterior Predictive Assessment of Model Fitness via Realized Discrepancies." *Statistica Sinica* 6: 733–807.
- Hartig, F. (2019). *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level/Mixed) Regression Models*. R package version 0.2.4.
- Hartzel, J., A. Agresti, and B. Caffo. 2001. "Multinomial Logit Random Effects Models." *Statistical Modelling* 1: 81–102.
- Hedeker, D. 2003. "A Mixed-Effects Multinomial Logistic Regression Model." *Statistics in Medicine* 22: 1433–46.
- Holmes, C. C., and L. Held. 2006. "Bayesian Auxiliary Variable Models for Binary and Multinomial Regression." *Bayesian Analysis* 1 (1): 145–68.
- Larson, W. 2015. *2014 Projection Review*. (Updated). <https://community.Qfangraphs.com/2014-projection-review-updated/> (accessed June 05, 2019).
- Lindsey, G. R. 1959. "Statistical Data Useful for the Operation of a Baseball Team." *Operations Research* 7 (2): 197–207.
- Mahalanobis, P. C. 1936. "On the Generalised Distance in Statistics." *Proceedings of the National Institute of Sciences of India* 2 (1): 49–55.
- Null, B. 2009. "Modeling Baseball Player Ability with a Nested Dirichlet Distribution." *Journal of Quantitative Analysis in Sports* 5 (2), <https://doi.org/10.2202/1559-0410.1175>.
- Pettitt, A. N., T. T. Tran, M. A. Haynes, and J. Hay. 2006. "A Bayesian Hierarchical Model for Categorical Longitudinal Data from a Social Survey of Immigrants." *Journal of the Royal Statistical Society: Series A* 169 (1): 97–114.
- Polson, N. G., J. G. Scott, and J. Windle. 2013. "Bayesian Inference for Logistic Models Using Polya-Gamma Latent Variables." *Journal of the American Statistical Association* 108: 1339–49.

- Powers, S., T. Hastie, and R. Tibshirani. 2018. "Nuclear Penalized Multinomial Regression with an Application to Predicting at Bat Outcomes in Baseball." *Statistical Modeling* 18 (5–6): 388–410.
- Quesenberry, C. P., and D. C. Hurst. 1964. "Large Sample Simultaneous Confidence Intervals for Multinomial Proportions." *Technometrics* 6 (2): 191–5.
- Rosner, B., F. Mosteller, and C. Youtz. 1996. "Modeling Pitcher Performance and the Distribution of Runs Per Inning in Major League Baseball." *The American Statistician* 50 (4): 352–60.
- Rubin, D. 1984. "Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician." *Annals of Statistics* 12: 1151–72.
- Sison, C. P., and J. Glaz. 1995. "Simultaneous Confidence Intervals and Sample Size Determination for Multinomial Proportions." *Journal of the American Statistical Association* 90 (429): 366–9.
- Smithson, M., and J. Verkuilen. 2006. "A Better Lemon Squeezer? Maximum-Likelihood Regression with Beta-Distributed Dependent Variables." *Psychological Methods* 11 (1): 54–71.
- Tango, T. 2004. *Tango on Baseball*. <http://www.tangotiger.net/archives/stud0346.shtml> (accessed April 01, 2019).
- Tango, T. 2012. *Marcel 2012*. <http://www.tangotiger.net/marcel/> (Accessed April 01, 2019).
- Wickramasinghe, I. 2014. "Predicting the Performance of Batsmen in Test Cricket." *Journal of Human Sport and Exercise* 9 (4): 744–51.