Predicting Rookie Batting Value in Major League Baseball

Paul Brendel

Determining whether a minor league batter will be successful at the major league level is one of the most important decisions a Major League Baseball (MLB) franchise makes every year in regards to the team's farm system. One way to measure an individual player's batting value is through a statistic called Batting Runs Above Average (BRAA). BRAA represents the number of runs above or below average a player has added as a hitter, adjusted for league and home park.[1] BRAA is designed to be Normally distributed around a mean of 0, representing a player who provides average batting value. The ability to predict major league BRAA for rookie call-ups would be very valuable information for a team because it could help dictate whether or not a player should be called up or when a player should be called up to the majors. This analysis attempts to predict rookie BRAA using a combination of minor league statistics and prospect grades via a Bayesian model.

Seven total predictors will be used to predict BRAA: 5 are minor league statistics, and 2 are scouting grades. The minor league statistics came from the most recent league and season in which that player recorded 100 plate appearances (PA). Weighted On-Base Average (wOBA) **(1)** is an advanced statistic that "combines all the different aspects of hitting into one metric, weighting each of them in proportion to their actual run value."[1] Weighted On-Base Average should be a better predictor of BRAA than slugging percentage (SLG) or on-base percentage (OBP) since each hitting outcome incorporated in wOBA is weighted based on run value. SLG and OBP were not included in the analysis due to multicollinearity problems. The formulas for BRAA, wOBA, SLG, and OBP are included in Table 1 of the Appendix. For the purposes of this analysis, wOBA was multiplied by 1000. Walk percentage (BB%) **(2)** indicates how often a player walks per plate appearance, and strikeout percentage (K%) **(3)** measures how often a player strikes out per plate appearance. BB% and K% are general measures of plate discipline and contact skills. The ideal batter would have a high walk rate and a low strikeout rate. Interactions between wOBA and BB% **(4)** and wOBA and K% **(5)** were also analyzed.

The prospect evaluators at MLB.com give prospects a subjective score between 20 and 80 in regards to Hitting **(6)** and Power **(7)**.[2] These measures roughly translate to a player's ability to reach base and produce extra-base hits, respectively. A grade of 50 represents major league average and each 10 point increment represents a standard deviation better or worse than average.[3] The distribution of each grade should be close to Normally distributed.[3] These grades were only given to the top 20 prospects on each team before the season started and since some rookies in 2015 were not regarded as top-20 prospects before the season our data set has missing values for prospect grades for 17 players.

During the 2015 MLB regular season, 53 rookies had at least 200 PA. Two rookies (Yasmany Tomas and Jung-ho Kang) are not included in the analysis because they were recruited from international leagues. The remaining 51 players were analyzed in this study. These rookies had a mean BRAA of 0.12 with a standard deviation of 10.11. The first 10 observations can be seen in Table 2 of the Appendix. BRAA and minor league statistics were obtained from fangraphs.com and prospect grades came from mlb.mlb.com/mlb/prospects.

As can be seen in Table 1 of the Appendix, BRAA is influenced by the number of plate appearances during the corresponding season. The rookies in this analysis were called up at varying points throughout the season and thus recorded different numbers of PA (with a minimum of 200 PA). To ensure that there was no correlation between BRAA and PA, these variables were plotted against each other in a scatter plot (Figure 1). The randomness in this plot and the Pearson correlation coefficient of 0.13 indicates a lack of correlation between these two variables.

Since the rookies in this analysis were called up from either the AAA or AA division, mean BRAA was calculated based on whether the player's last 100 minor league plate appearances came from AAA or AA and were from 2015 or 2014 (Table 3). Each subgroup had a mean BRAA near that of the overall mean BRAA with the exception of the 14.8 mean BRAA of players from AA, 2015. However, this subgroup had a sample size of only 2 players, preventing any real conclusions to be drawn.

The following model using multiple imputation was ran in R to obtain posterior estimates:

$$Y_i \mid \text{mu}_i, \text{tauy} \sim N(\text{mu}_i, \text{tauy})$$
$$\text{mu}_i = \text{beta0} + \text{beta1}*\text{wOBA}_i + \text{beta2}*\text{BB\%}_i + \text{beta3}*\text{K\%}_i + \text{beta4}*\text{wOBA}*\text{BB\%}_i + \text{beta5}*\text{wOBA}*\text{K\%}_i + \text{beta6}*\text{Hitting}_i + \text{beta7}*\text{Power}_i$$
$$\text{beta}_j \sim N(\text{my}_j, \text{vary}_j)$$

$$\text{Power}_i \mid \text{mux7}_i, \text{taux7} \sim N(\text{mux7}_i, \text{taux7})$$
$$\text{mux7}_i = \text{delta0} + \text{delta1}*\text{wOBA}_i + \text{delta2}*\text{BB\%}_i + \text{delta3}*\text{K\%}_i + \text{delta4}*\text{wOBA}*\text{BB\%}_i + \text{delta5}*\text{wOBA}*\text{K\%}_i + \text{delta6}*\text{Hitting}_i$$

$$\text{delta}_{jj} \sim N(\text{mx7}_{jj}, \text{varx7}_{jj})$$

$$\text{Hitting}_i \mid \text{mux6}_i\,, \text{taux6} \sim N(\text{mux6}_i, \text{taux6})$$
$$\text{mux6}_i = \text{gamma0} + \text{gamma1}*\text{wOBA}_i + \text{gamma2}*\text{BB\%}_i + \text{gamma3}*\text{K\%}_i + \text{gamma4}*\text{wOBA}*\text{BB\%}_i + \text{gamma5}*\text{wOBA}*\text{K\%}_i$$
$$\text{gamma}_{jjj} \sim N(\text{mx6}_{jjj}, \text{varx6}_{jjj})$$

$$\text{sigmay} \sim \text{Gamma}(\text{sy.a}, \text{sy.b});\ \text{tauy} = \text{sigmay}^2$$
$$\text{sigmax7} \sim \text{Gamma}(\text{sx7.a}, \text{sx7.b});\ \text{taux7} = \text{sigmaxy}^2$$
$$\text{sigmax6} \sim \text{Gamma}(\text{sx6.a}, \text{sx6.b});\ \text{taux6} = \text{sigmax6}^2$$

The fangraphs.com article describing wOBA was used to obtain the mean for the wOBA*1000 prior.[1] According to this article: "A good rule of thumb is that 20 points of wOBA is worth about 10 runs above average per 600 PA." Thus, our mean prior for wOBA*1000 was set to 2, corresponding to the increase in wOBA*1000 that leads to a 1 unit increase in BRAA.

The expert opinion of the investigator was used to obtain the mean and standard deviations for all other priors. This elicitation began with the generation of a point estimate. To determine the investigator's uncertainty in the estimate, he was also asked to provide a 99% confidence interval. Since experts generally give overly optimistic estimations when specifying intervals with high probability content, these 99% confidence intervals were treated as 95% confidence intervals, as seen in Bedrick et al.[4] For the wOBA*1000 prior mentioned above, the investigator believed that the given mean had a 99% confidence interval between 1.5 and 2.5. The prior standard deviation was therefore determined by dividing this interval by 4 (corresponding to a 95% confidence interval), resulting in a value of 0.25.

The investigator believed that a .75 unit increase in BB% led to a 1 unit increase in BRAA with a 99% confidence interval of (0, 1.5). Converting this to a 95% confidence interval resulted in a prior standard deviation of 0.375. This procedure was repeated for the rest of the priors used to predict BRAA (K%, Hitting, Power). The prior for K% was estimated to have a particularly wide confidence interval because some players have proven the ability to be very successful batters despite high strikeout rates due to an ability to minimize the number of outs from balls in play (i.e. Chris Davis). The priors for the interaction terms (wOBA*BB% and wOBA*K%) were set to values that would give them essentially no influence on the posterior estimates (mean 0, standard deviation 10) because the investigator had no knowledge of the effect of these terms on BRAA.

The intercept in this analysis corresponds to a player with a value of 0 for each of the predictors since the predictors are interpreted on a straight scale and not a mean-centered scale. Since the scenario of a player having a predictor value of 0 is difficult to comprehend in regards to the resulting BRAA, the investigator can only assume that the intercept value should be negative. A linear regression was conducted using all the predictors in order to get a "ballpark" estimate of the intercept. Using this information, the prior intercept was set to -50 and given a standard deviation of 10.

The prior for the BRAA standard deviation (sigmay) was given a point estimate of 10 by the investigator. The 99% confidence interval elicited for this estimate was (5,15) so this range was treated as a 95% confidence interval and divided by 4, resulting in 2.5 as the standard deviation of the BRAA standard deviation. To obtain the necessary values of a and b to use in the prior Gamma distribution for the standard deviation of BRAA, the following equations were used: $a/b = 10$; $a/b^2 = 2.5^2$. Solving the equations resulted in final values of a=16 and b=1.6. The values for the prior estimates used to predict BRAA can be seen in Table 4 and the resulting posterior estimates can be seen in Table 5.

The investigator decided to allow for the prediction of the 17 missing values for Hitting and Power to be determined almost entirely by the data. Therefore, most of the priors used for predicting the missing values of Hitting and Power were given values that would result in unspecific priors and data-driven posteriors. The investigator allowed the same wOBA prior mean and variance used in predicting BRAA to be used in the prediction of the Hitting and Power missing values since these statistics operate on nearly the same scale as BRAA and should be equally impacted by wOBA (1.5 was multiplied to the variance to account for the added uncertainty).

The posterior estimates suggest that BB% and Hitting grade were the only significant predictors of BRAA. Increasing minor league BB% by about 7.7% and increasing the prospect hitting grade by about 4.1 points should lead to a 10 unit increase in BRAA. The estimate of K% was too varied to be a significant predictor of BRAA. K% may be useful to explain what kind of approach a hitter has at the plate (i.e. how often a player swings at pitches), but it doesn't seem to be very predictive of actual hitting success. Hitting was a slightly better predictor of BRAA than Power perhaps because batting skills associated with reaching base may transition better to major league pitching compared to batting for power. Major league pitchers might be more adept at "pitching around" power hitters. For example, prospect Joey Gallo had a perfect (80) power grade going into the 2015 season, but he wasn't able to stay on the major league roster for very long because he was not able to generate many runs nor reach base at an effective rate. Figure 2 shows comparisons between the posteriors of BB% vs. K% and Hitting vs. Power. Based on the Hitting and Power values generated for Odubel Herrera, the model seems sufficient at replacing the missing values for Hitting and Power. There does not seem to be any interaction effect between wOBA and BB% nor wOBA and K%—the 95% confidence interval of each estimate includes both positive and negative numbers. When major league teams are evaluating the potential batting ability of prospects, both minor league statistics and prospect grades should be used in consideration.

To assess for any potential convergence problems, an autocorrelation plot and time series plot was created for the posterior estimates of wOBA and BB% (Figures 3 and 4). Each autocorrelation plot indicates that the time series for wOBA and BB% is random — there is no autocorrelation between adjacent nor non-adjacent observations. Each time series plot shows that the mean estimates of wOBA and BB% are consistently around .06 and .8, respectively, throughout all 10,000 iterations. Based off these results from the wOBA and BB% coefficients, it is assumed that the model and resulting estimates do not have any problems with convergence.

A sensitivity analysis was performed by comparing the original posterior results to those from a model in which the variance for each prior in the linear predictor of BRAA was divided by 2 or multiplied by 2. Overall, these changes to the variance did not result in major changes to the posterior estimates. After decreasing the variance of each prior, wOBA became a significant predictor of BRAA; however, its effect on BRAA is still minimal. After increasing the variance of each prior, the posterior estimate of BB % remained the same, but was no longer a significant predictor of BRAA suggesting that it might not be a very robust predictor of BRAA. The prior estimate of BB% may have been adding some certainty to a predictor whose data was otherwise quite noisy. The sensitivity analysis does not change the overall conclusion that BB% and Hitting grade were the best predictors of BRAA. Future analyses should consider consulting more experienced baseball experts in determining the prior estimates. The JAGS program used in R statistical software for this investigation is included at the end of the Appendix.

Appendix

Statistic	Formula
BRAA	wRAA + (lgR/PA-(PF*lgR/PA))*PA + (lgR/PA-(AL or NL non-pitcher wRC/PA))*PA
wOBA*	[(.69*uBB)+(.72*HBP)+(.88*1B)+(1.26*2B)+(1.59*3B)+(2.06*HR)]/(AB+BB-IBB+SF+HBP)
SLG	[(1B)+(2*2B)+(3*3B)+(4*HR)]/AB
OBP	(H+BB+HBP)/(AB+BB+HBP+SF)

Table 1. Formulas for BRAA, wOBA, SLG, and OBP. *These weights, which change slightly from year to year, are specific for 2015.

Player	BRAA	wOBA*1000	BB%	K%	Hitting	Power
Kris Bryant	27.2	439	14.5	28.6	55	75
Matt Duffy	11.6	379	10.1	15.8	55	30
Francisco Lindor	14.4	347	9.5	14.5	60	40
Odubel Herrera	6.3	356	7.1	17.2	NA	NA
Carlos Correa	16.3	347	10.6	12.6	60	70
Randal Grichuk	15.2	345	5.9	22.9	50	55
Addison Russell	-6.1	388	4.4	17.1	60	60
Joc Pederson	10.5	439	18.1	26.9	55	55
Billy Burns	1.6	349	8.9	16.8	50	20
Devon Travis	9.7	365	8.4	13.6	60	40

Table 2. First 10 observations of data set.

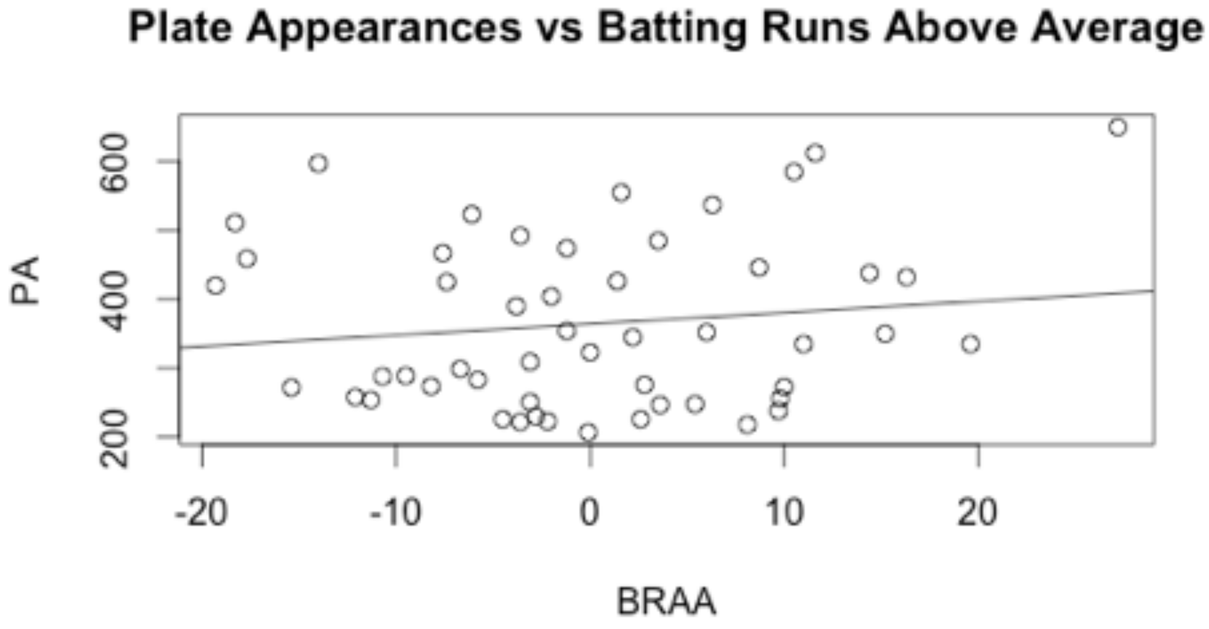

Figure 1. 2015 Rookie Major League Plate Appearances (minimum 200 PA) vs. Batting Runs Above Average.

	n	BRAA
AAA, 2015	18	0.62 (8.98)
AAA, 2014*	21	-1.03 (11.14)
AA, 2015	2	14.8 (6.79)
AA, 2014	10	-1.27 (8.98)

Table 3. Batting Runs Above Average based on the most recent minor league division and year that the player recorded at least 100 PA. *Includes one player from 2013 due to injuries in 2014.

Prior	Mean	Standard deviation	Variance
Intercept	-50	10	100
wOBA*1000	2	0.25	0.063
BB%	0.75	0.375	0.141
K%	-1.5	2	4
(wOBA*1000)*BB%	0	10	100
(wOBA*1000)*K%	0	10	100
Hitting	1	0.5	0.25
Power	1	0.5	0.25
Sigmay	10	2.5	6.25

Table 4. Priors used in the prediction of BRAA.

Posterior	Mean	Standard deviation	2.5 percentile	97.5 percentile
Intercept	-54.07	9.58	-73.49	-35.87
wOBA*1000	0.062	0.044	-0.026	0.14
BB%	0.77	0.37	0.052	1.52
K%	0.43	0.84	-1.33	2.03
(wOBA*1000)*BB%	-0.00078	0.0015	-0.0037	0.0021
(wOBA*1000)*K%	-0.0013	0.0022	-0.0056	0.0028
Hitting	0.41	0.20	0.033	0.83
Power	0.21	0.11	-0.012	0.43
Sigmay	9.02	0.95	7.35	11.07
Hitting for Odubel Herrera	52.94	6.50	40.01	65.81
Power for Odubel Herrera	46.04	11.23	23.90	67.86

Table 5. Posterior estimates of the predictors used to predict BRAA and for the missing values of Hitting and Power for Odubel Herrera.

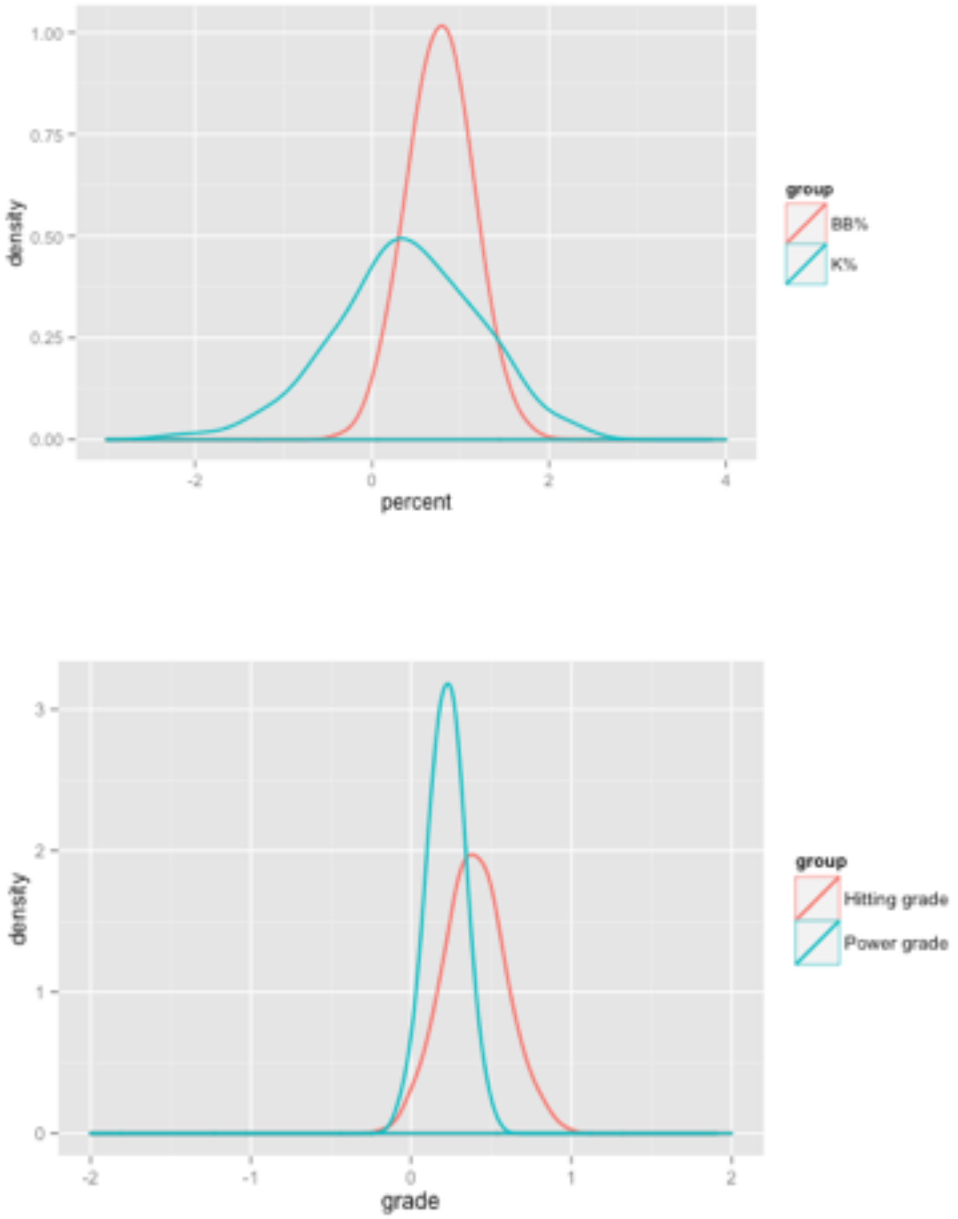

Figure 2. Posterior estimates of BB% vs. K% and Hitting vs. Power

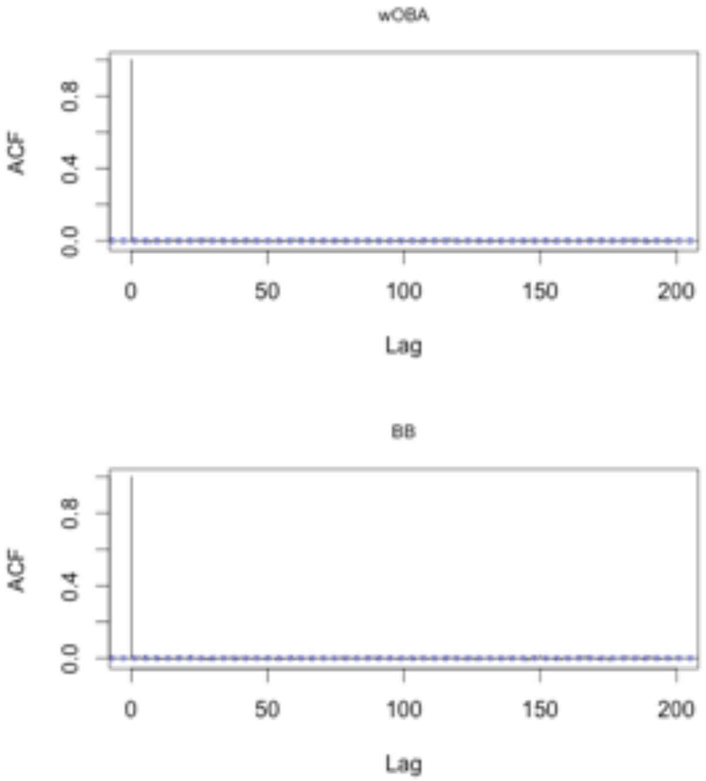

Figure 3. Autocorrelation plots of the posterior estimates of wOBA and BB%.

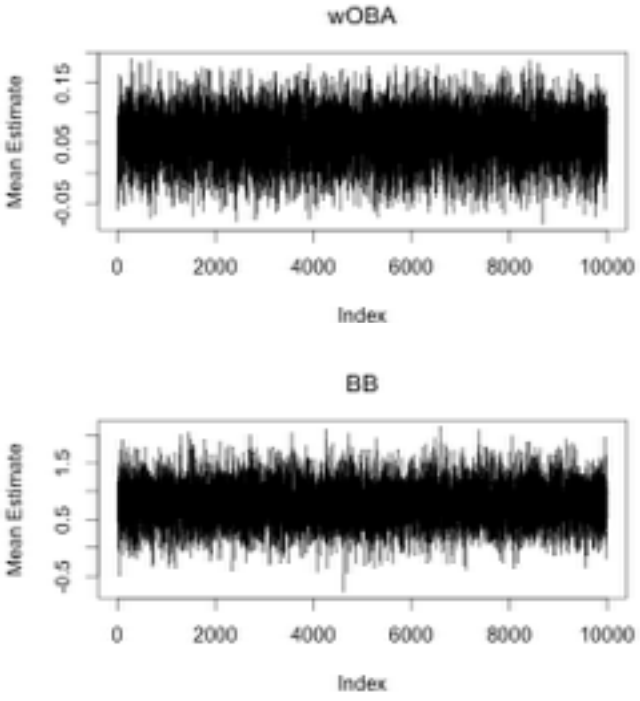

Figure 4. Time series plots of the posterior estimates of wOBA and BB%.

	Low Variance (Variance/2)	Normal	High Variance (Variance*2)
Intercept	-56.13 (-69.61, -42.14)	-54.07 (-73.49, -35.87)	-56.01 (-83.24, -33.16)
wOBA*1000	0.083 (0.011, 0.16)*	0.062 (-0.026, 0.14)	0.057 (-0.032, 0.16)
BB%	0.77 (.27, 1.29)*	0.77 (0.052, 1.52)*	0.77 (-0.25, 1.80)
K%	0.26 (-1.20, 1.90)	0.43 (-1.33, 2.03)	0.60 (-1.25, 2.50)
(wOBA*1000)*BB%	-0.0011 (-0.0039, 0.0016)	-0.00078 (-0.0037, 0.0021)	-0.00056 (-0.0042, 0.0030)
(wOBA*1000)*BB%	-0.0016 (-0.0060, 0.0026)	-0.0013 (-0.0056, 0.0028)	-0.0019 (-0.0073, 0.0031)
Hitting	0.35 (-0.015, 0.70)	0.41 (0.033, 0.83)*	0.42 (0.0053, 0.83)*
Power	0.22 (-0.014, 0.46)	0.21 (-0.012, 0.43)	0.21 (-0.04, 0.44)
Sigmay	9.20 (7.51, 11.37)	9.02 (7.35, 11.07)	9.00 (7.34, 11.04)

Table 6. Results of sensitivity analysis based on changing the variance of the priors in the linear predictor of BRAA. *Significant predictor of BRAA

JAGS Output:

```
sink("Finalmodel.txt")
cat("
    model
    {
    for(i in 1:N) {
    y[i] ~ dnorm( muy[i], tauy)
    muy[i] <- inprod(betay[1:8], x[i,1:8])

    x[i,8] ~ dnorm(mux7[i], taux7)
    mux7[i] <- inprod(beta7[1:7], x[i,1:7])
    x[i,7] ~ dnorm(mux6[i], taux6)
    mux6[i] <- inprod(beta6[1:6], x[i,1:6])
    }
    for (j in 1:K) {
    betay[j] ~ dnorm(my[j], precy[j] )
    precy[j]<- 1/vary[j]
    }
    for (jj in 1:(K-1)) {
    beta7[jj] ~ dnorm(mx7[jj], precx7[jj])
    precx7[jj]<- 1/varx7[jj]
    }
    for (jjj in 1:(K-2)) {
    beta6[jjj] ~ dnorm(mx6[jjj], precx6[jjj] )
    precx6[jjj]<- 1/varx6[jjj]
    }
    tauy <- 1/sigmay^2
    taux7 <- 1/sigmax7^2
    taux6 <- 1/sigmax6^2
    sigmay ~ dgamma(sy.a,sy.b)
    sigmax7 ~ dgamma(sx7.a,sx7.b)
    sigmax6 ~ dgamma(sx6.a,sx6.b)
    }

    ",fill = TRUE)
sink()

###

rawxyinits = matrix(data=scan(), byrow=T, ncol=7)
NA  NA  NA  NA  NA  NA  NA
NA  NA  NA  NA  NA  NA  NA
NA  NA  NA  NA  NA  NA  NA
NA  NA  NA  NA  NA  50  50
NA  NA  NA  NA  NA  NA  NA
NA  NA  NA  NA  NA  NA  NA
NA  NA  NA  NA  NA  NA  NA
NA  NA  NA  NA  NA  NA  NA
NA  NA  NA  NA  NA  NA  NA
NA  NA  NA  NA  NA  NA  NA
NA  NA  NA  NA  NA  NA  NA
NA  NA  NA  NA  NA  NA  NA
NA  NA  NA  NA  NA  NA  NA
NA  NA  NA  NA  NA  NA  NA
NA  NA  NA  NA  NA  NA  NA
NA  NA  NA  NA  NA  NA  NA
NA  NA  NA  NA  NA  NA  NA
NA  NA  NA  NA  NA  50  50
NA  NA  NA  NA  NA  50  50
NA  NA  NA  NA  NA  NA  NA
NA  NA  NA  NA  NA  NA  NA
NA  NA  NA  NA  NA  NA  NA
NA  NA  NA  NA  NA  NA  NA
NA  NA  NA  NA  NA  NA  NA
NA  NA  NA  NA  NA  50  50
NA  NA  NA  NA  NA  NA  NA
NA  NA  NA  NA  NA  NA  NA
NA  NA  NA  NA  NA  50  50
NA  NA  NA  NA  NA  NA  NA
NA  NA  NA  NA  NA  50  50
NA  NA  NA  NA  NA  NA  NA
NA  NA  NA  NA  NA  NA  NA
NA  NA  NA  NA  NA  50  50
NA  NA  NA  NA  NA  50  50
NA  NA  NA  NA  NA  50  50
NA  NA  NA  NA  NA  50  50
NA  NA  NA  NA  NA  NA  NA
NA  NA  NA  NA  NA  50  50
NA  NA  NA  NA  NA  NA  NA
NA  NA  NA  NA  NA  50  50
NA  NA  NA  NA  NA  NA  NA
NA  NA  NA  NA  NA  50  50
NA  NA  NA  NA  NA  NA  NA
NA  NA  NA  NA  NA  50  50
NA  NA  NA  NA  NA  NA  NA
NA  NA  NA  NA  NA  NA  NA
NA  NA  NA  NA  NA  50  50
NA  NA  NA  NA  NA  50  50
NA  NA  NA  NA  NA  NA  NA
NA  NA  NA  NA  NA  50  50
NA  NA  NA  NA  NA  NA  NA

xmat = cbind(1,Final.Data[,3:9])
colnames(xmat)[1] = "Intercept"

Data = list(N = 51, K = 8, my = c(-50, 2, 0, 0, .75, -1.5, 1, 1),
            vary = c(100, .0625, 100, 100, .141, 4, .25, .25),
            mx7 = c(0, 2, 0, 0, 0, 0, 0),
            varx7 = c(100, .0625*1.5, 100, 100, 100, 100, 100),
            mx6 = c(0, 2, 0, 0, .75, -1.5),
            varx6 = c(100, .0625*1.5, 100, 100, 100, 100),
            sy.a  = 16, sy.b  = 1.6,
            sx7.a = 1, sx7.b = 1,
            sx6.a = 1, sx6.b = 1,
            x = xmat, y = Final.Data[,2])

Finalinits = list(betay = c(-50, 2, 0, 0, 0, 0, 0, 0), sigmay = 10,
             beta7 = c(0, 2, 0, 0, 0, 0, 0), sigmax7 = 10,
             beta6 = c(0, 2, 0, 0, 0, 0), sigmax6 = 10,
             x = cbind(NA, rawxyinits))

Inits = rep(list(Finalinits),3)

Parameters = c("betay", "sigmay", "sigmax7", "sigmax6", "x[4,7]", "x[4,8]")
```

```
proc.time()
run1 = jags(Data, Inits, Parameters, "Finalmodel.txt",
          n.chains=3, n.iter=11000, n.burnin=1000, n.thin=1)
proc.time()
```

References

1. Complete List (Offense). Fangraphs. Dec., 2015. http://www.fangraphs.com/library/offense/offensive-statistics-list/

2. 2014 Prospect Watch. MLB.com. http://mlb.mlb.com/mlb/prospects/watch/y2014/#list=prospects

3. McDaniel K.: Scouting Explained: The 20-80 Scouting Scale. Fangraphs. Sep., 2014. http://www.fangraphs.com/blogs/scouting-explained-the-20-80-scouting-scale/

4. Bedrick EJ, Christensen R, Johnson W.: Bayesian Binomial Regression: Predicting Survival at a Trauma Center. *The American Statistician*, Vol. 51, No. 3. (Aug., 1997), pp. 211-218.