True wOBA: Estimation of true talent level for batters

Scott Powers and Eli Shayer

February 26, 2016

1 Introduction

We begin with a thought exercise. Which statistic is more volatile: on-base percentage (OBP) or batting average on balls in play (BABIP)? To sharpen the question mathematically, suppose we observe the following data for batter B:

- OBP_1, his OBP in his first 150 plate appearances;
- $BABIP_1$, his BABIP on his first 150 *balls in play*;
- OBP_2, his OBP in his second 150 plate appearances; and
- $BABIP_2$, his BABIP on his second 150 balls in play.

Which do you expect to be closer: OBP_1 to OBP_2, or $BABIP_1$ to $BABIP_2$? In other words, which is a better predictor of itself: OBP or BABIP? On the one hand, OBP is the statistic made famous by *Moneyball* for its merits in player evaluation. On the other hand we have the notoriously "luck-driven" statistic BABIP. We know that OBP stabilizes at 460 plate appearances while BABIP stabilizes at 820 balls in play (Carleton, 2012).

The answer? *On-base percentage is more volatile.* When using OBP_1 to predict OBP_2, the root mean square error (RMSE) is .062. When using $BABIP_1$ to predict $BABIP_2$, the RMSE is a hair smaller, at .058. Mathematically, this makes perfect sense. The league average BABIP is .299, and the league average OBP is .317. So the standard binomial assumption implies that OBP should have slightly higher variance than BABIP. Variance is, after all, another word for volatility. To reiterate, *on average, observed BABIP is just as close to true talent as observed OBP is, and possibly closer.*

Why, then, does BABIP get a bad reputation? It has to do with the definition of stabilization rate, which is based on the number of trials needed from each player in two different samples in order for the correlation between the two estimates of the statistic (one from each sample) to exceed some threshold. Under the assumption that observed OBP is equal to true talent plus noise, we have

$$OBP_1 = OBP_{\text{true}} + \epsilon_1, \qquad \text{and} \qquad OBP_2 = OBP_{\text{true}} + \epsilon_2,$$

where ϵ_1 and ϵ_2 are independent and identically distributed, say with variance $\sigma^2_{\epsilon,\text{OBP}}$. If the population variance in OBP true talent is $\sigma^2_{\text{pop,OBP}}$, then

$$\rho_{\text{OBP}} \equiv \text{Corr}(OBP_1, OBP_2) = \frac{\text{Cov}(OBP_1, OBP_2)}{\sqrt{\text{Var}(OBP_1)\text{Var}(OBP_2)}} = \frac{\sigma^2_{\text{pop,OBP}}}{\sigma^2_{\text{pop,OBP}} + \sigma^2_{\epsilon,\text{OBP}}}$$

Similarly,

$$\rho_{\text{BABIP}} \equiv \text{Corr}(BABIP_1, BABIP_2) = \frac{\sigma^2_{\text{pop,BABIP}}}{\sigma^2_{\text{pop,BABIP}} + \sigma^2_{\epsilon,\text{BABIP}}}.$$

We have observed that $\sigma^2_{\epsilon,\mathrm{OBP}}$ is slightly larger than $\sigma^2_{\epsilon,\mathrm{BABIP}}$, so if $\rho_{\mathrm{OBP}} > \rho_{\mathrm{BABIP}}$, then it must be the case that $\sigma^2_{\mathrm{pop,OBP}} > \sigma^2_{\mathrm{pop,BABIP}}$. Sure enough, this is exactly the case. Our estimate of the population variance of OBP skill is about 33% greater than our estimate of the population variance of BABIP skill. So the slow stabilization of BABIP is not driven by the volatility of the statistic but rather by the lack of spread between players in the distribution of underlying true talent. This is consistent with the interpretation that BABIP has more to do with luck than does OBP, because it means that there is less spread between good BABIP batters and bad BABIP batters than there is between good OBP batters and bad OBP batters.

Because the spread in true talent BABIP is less than the spread in true talent OBP, we can leverage this information to make even more accurate estimates of BABIP using the technique of regression to the mean. We leave the details of our implementation of regression to the mean to Appendix A of Tango et al. (2007), but it is a two-step procedure. In the first step, we estimate the underlying true talent population variance of the statistic by comparing the observed variance in the data to the theoretical variance if all players had equal skill. In the second step, we take for each player a weighted average of the population mean and his observed statistic, weighted by the inverse of the population variance and the inverse of the statistic's sample variance, respectively. This shrinkage of the estimates improves the RMSE of the OBP estimator from .062 to .049 and improves the RMSE of the BABIP estimator from .058 to .047. An important observation here is that correlation is not a good measure of the accuracy of a prediction because of its scale invariance. For example, the correlation of both the unregressed estimate of OBP and the regressed estimate of OBP with the observed OBP in the second sample is 0.23. The correlation is the exactly same for both estimators despite one being more accurate than the other. For this reason, throughout the paper we use mean square error (and RMSE) as our criterion for evaluating estimates.

Table 1: *Root mean square error for predicting the rate of each outcome using naive and regressed estimators. Each estimator uses the first 200 PAs for each batter to predict the rate at which that batter will produce the outcome in his second 200 PAs. The error is reported in percentage points, so for example the RMSE of 3.81 for the naive estimator of 1B rate corresponds to the difference between an estimate of 15.00% and an observation of 18.81%. The bottom row of the table gives the estimated population variance of the true talent for producing the corresponding outcome, in the same units (percentage points).*

	G	F	K	BB	HBP	1B	2B	3B	HR
Naive error	4.80	4.45	4.19	3.33	0.94	3.81	2.04	0.74	1.79
Regressed error	4.42	4.22	3.89	3.04	0.80	3.17	1.62	0.67	1.61
Population variance	15.85	20.13	29.10	6.26	0.24	7.02	0.45	0.13	1.88

We apply the same methodology to the rates of each of nine different outcomes of plate appearances, with the results presented in Table 1. Note that each outcome has a unique population variance, with smaller population variances necessitating more aggressive regression to the mean. Whereas strikeouts need very little regression to the mean, doubles and triples should be regressed aggressively.

We combine the regressed estimates summarized by Table 1 to obtain a regressed estimate of wOBA for each batter based on his first 200 PAs. We plot regressed wOBA versus observed wOBA in Figure 1. The plot illustrates how players with relatively few PAs are regressed further toward the mean than players with relatively many PAs. Also, the players with the most extreme observed wOBAs take larger steps toward the mean, reflecting that players with the best performances were likely lucky in addition to being good.

Note that because we have differentially regressed the different outcome types, less stable outcomes have been regressed more toward the mean than more stable outcomes. Players with a high BABIP, for example, will be expected to have a lower BABIP moving forward, according to this method. And in

Figure 1: *Players plotted by mean-regressed wOBA vs. observed wOBA. Players with fewer than 50 PAs are orange dots, and players with at least 50 PAs are blue circles. The horizontal dotted black line corresponds to league average wOBA, and the green diagonal line corresponds to True wOBA exactly equal to observed wOBA. We observe a steeper slope for the cloud of blue circles than for the cloud of orange dots because fewer PAs mean more aggressive regression to the mean.*

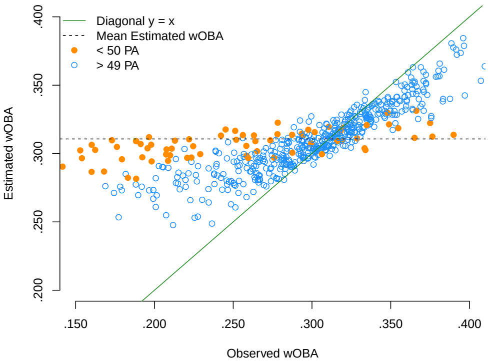
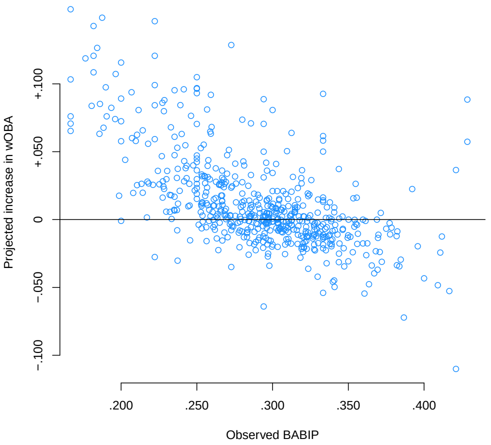


Figure 2: *Projected change in wOBA vs. observed BABIP. On the y-axis is regressed wOBA minus observed wOBA. The greater the BABIP, the lower the expected wOBA, relative to observed wOBA.*



fact, Figure 2 illustrates the relationship between observed BABIP and the difference between observed wOBA and regressed wOBA. The higher BABIP is, the lower regressed wOBA is relative to observed wOBA. This encapsulates and precisely quantifies the common understanding that players with high BABIP typically have been lucky and should expect their performance to drop.

Regression to the mean precisely quantifies claims that players with small numbers of PAs or high BABIPs are unlikely to sustain their performance. Moreover, it quantifies what performance we can expect moving forward. In Section 2 we connect regression to the mean with shrinkage estimators in linear regression, thus extending the approach to account for additional covariates and leading to the proposal of True wOBA. In Section 3 we present the results of True wOBA on the 2015 MLB regular season play-by-play data.

2 Methods

2.1 Data

The dataset used throughout this paper is the 2015 Major League Baseball regular season play-by-play dataset from Retrosheet. For each plate appearance the data include, among other things, the identity (B_i) of the batter, the identity (P_i) of the pitcher, the stadium (S_i) where the PA took place, an indicator (H_i) of whether the batter was on the home team and an indicator (O_i) of whether the batter had opposite handedness from the pitcher.

For the present work, we consider only plate appearances which result in one of the following outcomes: groundout (G), flyout (F), strikeout (K), base on balls (BB), hit by pitch (HBP), single (1B), double (2B), triple (3B) or home run (HR). If the batter reached base on an error, we count this as G or F depending on the trajectory of the batted ball. We count all fielder's choices as G. We discard intentional walks and catcher's interferences. We also discard PAs in which the batter is a pitcher. The result is a dataset of 176,560 PAs featuring 660 unique batters and 735 unique pitchers.

2.2 Regularization as regression to the mean

Suppose that batter X has skill β_X for hitting singles. Using Y to denote the outcome of a plate appearance for batter X, the probability model defined by

$$\mathbb{P}(Y = 1\text{B}) = \frac{e^{\alpha+\beta_X}}{1 + e^{\alpha+\beta_X}}$$

is a logistic regression model. The parameter α is an intercept term corresponding to the baseline likelihood of a single. The larger $\alpha + \beta_X$ is, the more likely it is that the PA results in a single.

More ambitiously, take B_i to be the identity of the batter in the i^{th} PA in the 2015 Retrosheet play-by-play data, for i from 1 to 176,560. A model which describes the distribution of all singles is given by:

$$\mathbb{P}(Y_i = 1\text{B}) = \frac{e^{\alpha+\beta_{B_i}}}{1 + e^{\alpha+\beta_{B_i}}}, \tag{1}$$

independently for each i. Note that, for fixed intercept α, this model is identical to one in which each batter X simply has some probability p_X of hitting a single. This is due to the one-to-one correspondence between β_X and p_X.

To fit the model (1), we seek to maximize its log-likelihood $\ell(\beta|B, Y)$:

$$\max_\beta \ell(\beta|B, Y) \equiv \max_\beta \sum_{i=1}^{176560} \left\{ Y_i \log\left(\frac{e^{\alpha+\beta_{B_i}}}{1 + e^{\alpha+\beta_{B_i}}}\right) + (1 - Y_i)\log\left(\frac{1}{1 + e^{\alpha+\beta_{B_i}}}\right) \right\} \tag{2}$$

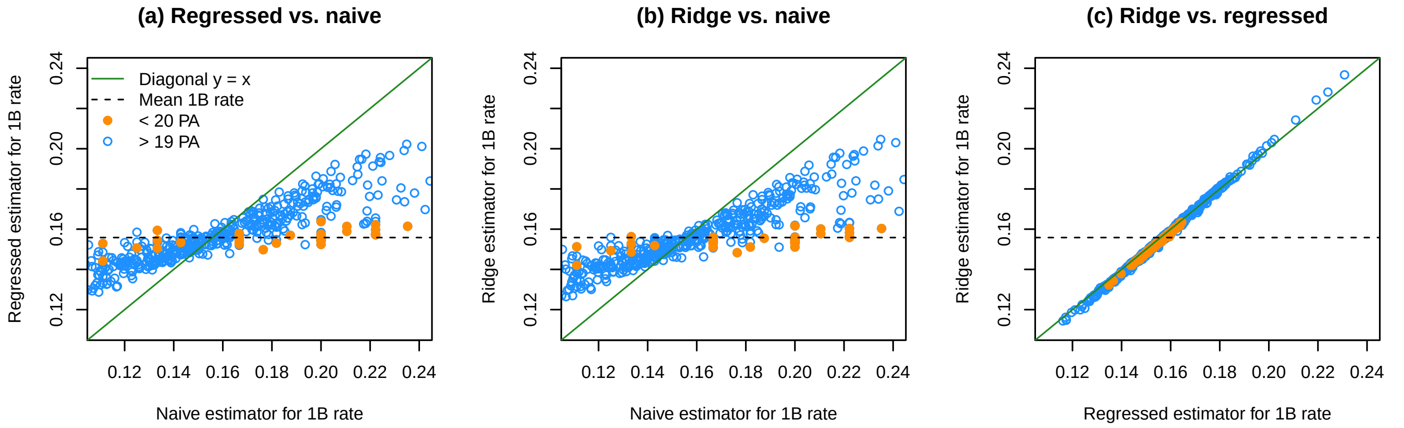
which turns out to be equivalent to using the observed fraction of singles for each batter as the estimate for his 1B rate. For example, a player who hit 15 singles in 100 PAs would be estimated to have a 1B rate of 15%. We call this the *naive* estimator of 1B rate. As discussed in Section 1, a batter estimator is the *regressed* estimator.

To relate our logistic regression model to regression to the mean, we appeal to the concept of *regularization* from machine learning. Instead of solving the optimization problem (2), instead consider

$$\max_\beta \ell(\beta|B, Y) - \lambda \sum_{X \in \mathcal{B}} \beta_X^2. \tag{3}$$

We have introduced a penaly parameter $\lambda > 0$, which is multiplied by the sum of the squares of the skills of our set $\mathcal{B}$ of batters. To maximize (3), we must choose β which explain the data well but are also not too large, because we pay a penalty for β which are far from zero. The result is that the skill estimates β_X are pulled toward zero, and thus the estimates p_X of 1B rate are pulled closer together, just as in regression to the mean. The optimization problem (3) is known as ridge logistic regression, and we call the resulting estimator of 1B rate the *ridge* estimator.

Figure 3: *Comparison of ridge estimator with regressed estimator of 1B rate. Batters with fewer than 20 PAs are orange dots, and batters with at least 20 PAs are blue circles. The horizontal dotted black line corresponds to league average 1B rate. The diagonal green line corresponds to where the two estimators are equal. Figure (a) plots regressed estimates against naive estimates while figure (b) plots ridge estimates against naive estimates. The similarity of figures (a) and (b) suggests the close correspondence of the ridge and the regressed estimators, which is confirmed by figure (c), which plots regressed estimates against ridge estimates.*



The regularization parameter λ is chosen through a process called cross validation. We split the training data into 10 folds and for each fold, we use the data not in that fold to predict the data in the fold, creating predictions for many different values of λ. The value of λ which leads to the lowest error overall in the cross validation step is the one we use to fit the model on the whole training set. Hence λ is chosen with the goal of minimizing prediction error in mind.

Table 2: *Root mean square error for predicting the rate of each outcome using naive, regressed and ridge estimators. Each estimator uses half of all plate appearance (selected at random) to predict the rate at which each batter will produce each outcome in the other half of the data. Error is evaluated on all players with at least 30 PAs in the held-out test set, and the error for each player is given weight proportional to his number of PAs in the held-out test set.*

Estimator	G	F	K	BB	HBP	1B	2B	3B	HR
Naive	4.41	4.45	4.25	2.60	1.04	3.66	2.21	0.82	1.71
Regressed	3.98	3.97	3.89	2.38	0.89	3.09	1.68	0.63	1.52
Ridge	3.97	3.99	3.90	2.39	0.88	3.08	1.67	0.64	1.51

Conceptually, the ridge estimator is similar to the regressed estimator because both apply a shrinkage to pull estimates of 1B rate toward a center. The similarity between the two estimators is illustrated by Figure 3, which shows that the two estimators give nearly identical results on the play-by-play data. Table 2 demonstrates that this relationship extends beyond only 1Bs to cover all of the other PA outcomes, as well. For each PA outcome, the ridge estimator leads to an improvement in prediction over the naive estimator, yielding similar results to the regressed estimator.

The key difference between the regressed estimator and the ridge estimator is how each method determines how aggressively to shrink the results toward the mean. The amount of shrinkage in regression to the mean is determined by the population variance parameter, which itself is estimated by comparing observed variance in the data to theoretical variance in binomial random variables, as described in Section 1. The amount of shrinkage for the ridge estimator, on the other hand, is determined by the regularization parameter λ. The larger it is, the more aggressively estimates are regressed toward their mean. Using cross validation to choose λ means that we are choosing the amount of shrinkage to minimize prediction error. This approach directly attacks our goal of minimizing prediction instead of estimating population variance as an intermediate step.

The bottom line is that instead of using regression to the mean, we can fit a ridge logistic regression model and get similar results. The advantage of this is that the logistic regression model can be generalized to simultaneously adjust for park effects, opponent quality and other factors. This upshot is the key idea behind our proposal, True wOBA.

2.3 True wOBA

For each outcome $k \in \{\text{G, F, K, BB, HBP, 1B, 2B, 3B, HR}\}$, the True wOBA model is

$$\mathbb{P}(Y_i = k) = \frac{e^{\alpha_k + \beta_{B_i k} + \gamma_{P_i k} + \delta_{S_i k} + \zeta_k H_i + \theta_k O_i}}{1 + e^{\alpha_k + \beta_{B_i k} + \gamma_{P_i k} + \delta_{S_i k} + \zeta_k H_i + \theta_k O_i}}. \tag{4}$$

As in Section 2.2, this is a logistic regression model for each outcome, but here it depends not only on the batter but also on the pitcher and other variables. Refer to Section 2.1 for a description of the data B_i, P_i, S_i, H_i and O_i. The fixed, unknown paramaters in (4) have the following interpretation:

- α_k: the baseline log-odds of outcome k;
- $\beta_{B_i k}$: the tendency of batter B_i to produce outcome k;
- $\gamma_{P_i k}$: the tendency of pitcher P_i to produce outcome k;
- $\delta_{S_i k}$: the tendency of stadium S_i to produce outcome k;
- ζ_k: the increase in log-odds of outcome k due to home field advantage; and
- θ_k: the increase in log-odds of k due to the batter having opposite handedness from the pitcher.

True wOBA models the likelihood of an outcome as a tradeoff between the batter's skill in producing the outcome and the pitcher's skill in preventing it, with park effects and other adjustments built in. To fit each logistic regression model we use ridge regression as in (3). Because of this ridge penalty, by construction the average batter has $\beta_{Bk} = 0$. Similarly, the average pitcher has $\gamma_{Pk} = 0$, and the average stadium has $\delta_{Sk} = 0$. We use the R package `glmnet` (Friedman et al., 2010) to fit the ridge logistic regression (with 10-fold cross validation to choose λ), yielding estimated coefficients α_k^*, ζ_k^*, θ_k^*, β_{Bk}^* for each batter B, γ_{Pk}^* for each pitcher P, and δ_{Sk}^* for each stadium S.

For each batter B, the corresponding estimate of the probability that he produces outcome k in a plate appearance against an average pitcher in an average stadium is given by

$$p_{Bk}^* = \frac{e^{\alpha_k^* + \beta_{Bk}^* + \zeta_k^*/2 + \theta_k^*/2}}{1 + e^{\alpha_k^* + \beta_{Bk}^* + \zeta_k^*/2 + \theta_k^*/2}}.$$

By using $\zeta_k^*/2$, this formulation mediates the two possibilities that the batter is home or away. Similarly, using $\theta_k^*/2$ mediates the two possibilities that the batter has opposite or same handedness as the pitcher.

Given p^*_{Bk} for each batter B and outcome k, we combine these estimates into True wOBA using the wOBA weights published by FanGraphs (FanGraphs, 2016).

For each pitcher P, the probability that he produced outcome k when facing an average batter in an average stadium is given by

$$p^*_{Pk} = \frac{e^{\alpha^*_k + \gamma^*_{Pk} + \zeta^*_k/2 + \theta^*_k/2}}{1 + e^{\alpha^*_k + \gamma^*_{Pk} + \zeta^*_k/2 + \theta^*_k/2}}.$$

We combine these estimated probabilities using the wOBA weights to get True wOBA against for pitchers.

2.4 Related work

Several recent papers (Brown, 2008; Null, 2009; Neal et al., 2010; Albert, 2015) have proposed methods for estimating or predicting batting statistics. In contrast with projection systems, these methods implicitly assume that players' skills remain constant between training sample and test sample (as opposed to modeling the aging of skills). True wOBA differs from the above by adjusting for park effects and quality of opposition.

A similar model which merits further discussion is Deserved Run Average (Judge and BP Stats Team, 2015, DRA). At the core of DRA is a regression model very similar to (4) which models the distribution of outcomes as a function of the tradeoff between the batter's and pitcher's skills, along with a plethora of control variables. This DRA model is much more complicated than True wOBA, including, for example, adjustments for the identity of the umpire, game time temperature and a home field advantage that varies by innning. Stripping the DRA model down to only the variables that True wOBA uses to facilitate comparison, a light version of this core piece of DRA models the wOBA on i^{th} plate appearance as

$$\text{wOBA}_i = \alpha + \beta_{B_i} + \gamma_{P_i} + \delta_{S_i} + \zeta_k H_i + \theta_k O_i + \epsilon_i, \text{ where}$$

$$\beta_B \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2_\beta), \qquad \gamma_P \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2_P), \qquad \text{and} \qquad \epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2_\epsilon). \tag{5}$$

So the batter skill paramaters β_B are not fixed but random, with a zero-mean normal distribution. Similarly, the pitcher skill parameters γ_P are random with a zero-mean normal distribution. The error terms ϵ_i are standard in linear regression. All other parameters, including the variance parameters σ^2_β, σ^2_P and σ^2_ϵ, are fixed and unknown. This is called a mixed effects linear model because it has both random and fixed parameters. It is fit via restricted maximum likelihood, which will we do using the R package `lme4` Bates et al. (2014).

By assuming that the β_B and the γ_P come from a mean-zero distribution, the mixed effects model (5) shrinks its estimates of the β_B and γ_P toward zero just as True wOBA does. This connection is discussed in more detail in Section 2.5. The key take-away from this section is that True wOBA is similar to a light version of the core component of DRA. We compare this light version of DRA to True wOBA in Section 3.1, under the name "mixed effects model."

2.5 Regularization vs. random effect modelling

We have two ways of achieving shrinkage of regression coefficient estimates: regularization as in (3) and random effect modelling as in (5). To facilitate comparison between the two approaches, we introduce another wOBA estimator here based just on the identity of the batter, as we did with the naive, regressed and ridge estimators from Section 2.2. We fit this model to each outcome category; for example

$$\mathbb{P}(Y_i = 1\text{B}) = \Phi(\alpha + \beta_{B_i}), \qquad \beta_B \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2_B), \tag{6}$$

where $\Phi(\cdot)$ is the standard normal cumulative density function. This is probit regression, an alternative to logistic regression. Once we fit this model using restricted maximum likelihood, for each batter B we have β^*_B, and the corresponding *random effect* estimator for the probability that B hits a single is $\Phi(\alpha^* + \beta^*_B)$. Table 3 compares the random effect estimator to the ridge estimator in terms of test root mean square error after fitting each on a randomly selected half of the play-by-play dataset. We observe that the two estimators yield virtually identical results.

Table 3: *Root mean square error for predicting the rate of each outcome using ridge and random effect estimators. Each estimator uses half of all plate appearance (selected at random) to predict the rate at which each batter will produce each outcome in the other half of the data. Error is evaluated on all players with at least 30 PAs in the held-out test set, and the error for each player is given weight proportional to his number of PAs in the held-out test set.*

Estimator	G	F	K	BB	HBP	1B	2B	3B	HR
Regressed	3.38	3.48	3.30	2.06	0.78	2.63	1.45	0.55	1.36
Random	3.38	3.49	3.35	2.06	0.77	2.64	1.44	0.56	1.35

The conceptual difference between the ridge estimator and the random effect estimator is similar to the difference between the ridge estimator and the regressed estimator. Like the regressed estimator, restricted maximum likelihood for the random effect estimator uses the variance observed in the response, relative to the expected variance if all players had the same abilities, to infer the population variance of player's abilities. In that sense the random effect model is a more direct extension of the regressed estimator to allow for a linear model with covariates.

The ridge estimator has the advantage over the random effect estimator that it could be further extended to a multinomial regression model, jointly modeling the outcome probabilities under the restriction that they must sum to one. However, as `glmnet` is currently set up, it does not allow for different penalties to be applied to the regression coefficients for different outcome categories in multinomial regression, making this extension an area for future work.

Computationally, the ridge estimator could be faster or slower than the random effect estimator, depending on the number of values of λ searched over in the cross validation step. In practice, we have not found a substantial difference between the two approaches of regularization and mixed effect modelling.

3 Results

3.1 Validation

We randomly selected some PAs to set aside as a test set and used the remaining PAs as a training set to fit four different wOBA estimators: naive wOBA, regressed wOBA, True wOBA and the mixed effects model. As an added wrinkle, the test set was not chosen uniformly at random. We biased the selection so that PAs which feature opposite-handed batters and pitchers have a 90% chance of being held out for the test set while PAs with same-handed batters and pitchers have a 10% chance of being held out for the test set. This change in circumstances corresponds to what a player might face if he plays against tougher competition or moves to a new home stadium. We expect True wOBA and the mixed effects model to pick up on the changing circumstances and make better predictions than regressed wOBA.

The result of our biased sampling scheme is a training set of 93,868 PAs and a test set of 82,692 PAs. We evaluate each method based on how its predicted wOBA compares with observed wOBA for the 349 batters who have more than 100 PAs in the test set. Table 4 summarizes the results. True

Table 4: *Out-of-sample prediction error for four wOBA estimators. The error reported is mean square error when predicting wOBA for each player with at least 100 PAs in the test set. The error for the prediction of each player is weighted according to how many PAs he has in the test set. The bottom row responds the standard error of our estimated mean square error. The error for True wOBA, in bold, is the lowest of all four.*

Estimator	Naive	Regressed	True	Mixed
Estimated MSE	0.00456	0.00220	**0.00173**	0.00180
Stadard error	±0.00044	±0.00018	±0.00014	±0.00015

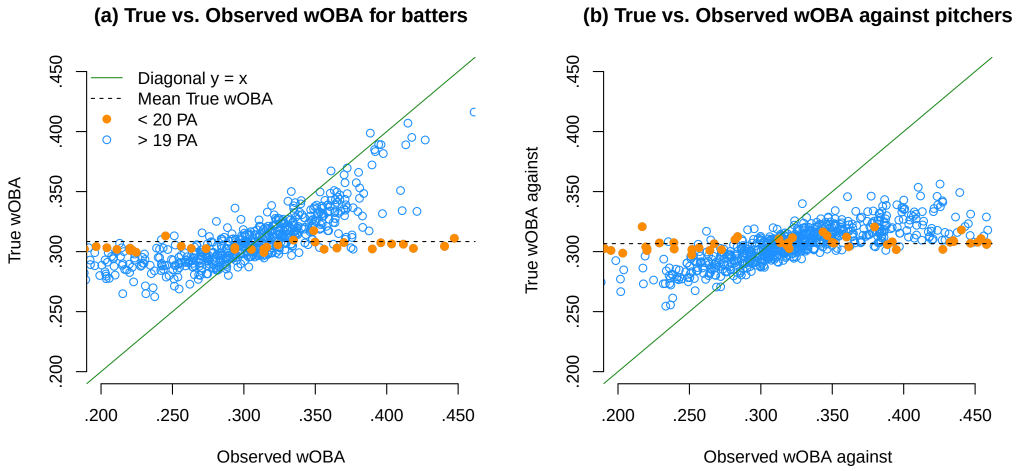
wOBA has the best performance, with strong evidence that it is better than regressed wOBA but only weak evidence that it is better than the mixed effects model.

We conclude that there is some value to True wOBA in terms of predicting future wOBA in a changing landscape. But the power of True wOBA is in its interpretation, which allows us to adjust performance for sample size and other variables.

3.2 True wOBA

We estimated True wOBA for all batters and True wOBA against all pitchers based on the full set of 176,560 plate appearances in the play-by-play dataset. Figure 4 shows the results. Players with fewer than 20 PAs tend to lie close to the horizontal line at mean wOBA because they have not demonstrated sufficient evidence that their true skill is any different from the average player. Players with the largest numbers of PAs tend to lie closer to the green diagonal line because there are sufficient data to more precisely estimate their true talent levels.

Figure 4: *Players plotted by True wOBA vs. observed wOBA. Figure (a) shows wOBA for all batters while figure (b) shows wOBA against all pitchers. Players with fewer than 20 PAs are orange dots, and players with at least 20 PAs are blue circles. The horizontal dotted black line corresponds to league average wOBA, and the green diagonal line corresponds to True wOBA exactly equal to observed wOBA.*



We observe a steeper slope for the batters' cloud of blue than for the pitchers' cloud of blue, reflecting more aggressive regression to the mean for pitchers. This dovetails with the common understanding in sabermetrics that pitchers have little control over the results of balls in play. In other words, the population variance in true talent level for outcomes on balls in play is relatively small for pitchers, so it makes sense to regress their results more aggressively.

Table 5: *Leaders and laggards. This table lists the top 5 and bottom 5 batters by True wOBA, as well as the top 5 and bottom 5 pitchers by True wOBA against.*

	Batter	Team	True wOBA	Pitcher	Team	True wOBA against
	Bryce Harper	WSN	.416	Jake Arrieta	CHC	.255
Top	Mike Trout	LAA	.407	Clayton Kershaw	LAD	.256
5	Jose Bautista	TOR	.399	Zach Greinke	LAD	.261
	Paul Goldschmidt	ARI	.395	Wade Davis	KCR	.267
	Joey Votto	CIN	.393	Dallas Keuchel	HOU	.267
	Alexi Amarista	SDP	.270	Jeremy Guthrie	KCR	.346
Bottom	Chris Owings	ARI	.269	Matt Boyd	DET	.346
5	René Rivera	TBR	.265	David Holmberg	CIN	.349
	Danny Santana	MIN	.265	Dustin McGowan	PHI	.354
	Omar Infante	KCR	.262	Allen Webster	ARI	.356

Table 5 shows the top and bottom players by True wOBA and True wOBA against. Most interestingly, consider the top three pitchers in MLB by True wOBA against: Jake Arrieta, Clayton Kershaw and Zach Greinke. These three were also the three candidates in a hotly contested 2015 NL Cy Young Award race. True wOBA agrees with the voters that Arrieta was the most deserving of the award, despite Kershaw being No. 1 by FIP-WAR and Greinke being No. 1 by RA9-WAR (FanGraphs, 2015). True wOBA mediates FIP-WAR and RA9-WAR by regressing results on balls in play aggressively (and optimally, in terms of out-of-sample prediction) but not ignoring them altogether.

The leaderboard also features 2015 AL Cy Young Award winner Dallas Keuchel at No. 5 by True wOBA allowed and 2015 NL MVP Bryce Harper at No. 1 by True wOBA. Mike Trout is a close second behind Harper at No. 2.

Table 6: *Largest differences between True wOBA and observed wOBA, among qualified players (min. 500 PAs). For each player ΔwOBA is equal to True wOBA minus observed wOBA. We list the top 5 gainers and top 5 losers among both batters and pitchers.*

	Batter	Team	ΔwOBA	Pitcher	Team	ΔwOBA against
	Wilson Ramos	WSN	+.022	Chris Rusin	COL	−.068
Top	Michael Taylor	WSN	+.021	Kyle Kendrick	COL	−.062
5	Albert Pujols	LAA	+.017	Jerome Williams	PHI	−.047
	Alcides Escobar	KCR	+.016	Matt Garza	MIL	−.045
	Chris Owings	ARI	+.014	Kyle Lohse	MIL	−.041
	Anthony Rizzo	CHC	−.035	Jacob deGrom	NYM	+.016
Bottom	Nolan Arenado	COL	−.037	Sonny Gray	OAK	+.016
5	Charlie Blackmon	COL	−.039	Clayton Kershaw	LAD	+.019
	Bryce Harper	WSN	−.045	Jake Arrieta	CHC	+.021
	David Peralta	ARI	−.046	Zach Greinke	LAD	+.023

Table 6 shows the players whose True wOBA is most higher than observed their wOBA and the players whose True wOBA is most lower than their observed wOBA. For example Wilson Ramos has a True wOBA which is 22 points higher than his observed wOBA. The primary drivers of the difference between True and observed wOBA are sample size (and fluctuations like in BABIP), park effects and quality of opposition. It is no surprise that Rockies pitchers Chris Rusin and Kyle Kendrick benefit the most from using True wOBA instead of observed wOBA in their evaluation. Similarly, Rockies batters Nolan Arenado and Charlie Blackmon are among those who drop the most going from observed wOBA to True wOBA. All of the pitchers in the bottom of Table 6 had stellar seasons, as did Harper. Their presence in the bottom of the table reflects the idea that those who performed best probably outperformed their true talent the most.

4 Discussion

Nominally, this is a paper introducing True wOBA, a wOBA estimator which simultaneously accounts for sample size, park effects and quality of opposition to facilitate comparisons between batters and between pitchers on a level playing field. But there is little thirst in the sabermetric literature for new batting metrics. The real contribution of this paper comes in three parts.

First, we argue against the use of stabilization rates when interpreting small sample sizes. For one matter, stabilization rates have more to do with the population variance in true skill level than variance in the observed statistic. For another, there is a spectrum of sample sizes, not just "too small" and "big enough," a nuance not addressed by stabilization rates. We advocate the use of regression to the mean to precisely quantify the uncertainty due to finite sample sizes. A feature of regression to the mean for wOBA is that it automatically learns that the population variance for underlying BABIP skill is relatively small, thus regressing BABIP more aggressively toward the mean. So regressed wOBA quantifies the assertion that we expect players with high BABIP to see their performance drop off more than players with low BABIP.

Second, we discuss in detail the relationship between regression to the mean and its extension to shrinkage regression estimators pioneered by Judge and BP Stats Team (2015). In particular, we relate the penalty parameter in regularized regression to the population variance parameter in regression to the mean to explore conceptually the fundamental difference between how much shrinkage is applied in each estimator.

Third, we make a novel comparison between regularized linear regression and random effects linear models. This comparison leads to a discussion of the relative strengths and weakness of the two approaches.

The goal of this paper is to review the fundamental statistical concepts that relate to evaluating players on a level playing field, accounting for sample size and other factors. No statistic readily available from FanGraphs.com or Baseball-Reference.com accomplishes this task, which should be the first step in player evaluation, as opposed to squinting at sample sizes and BABIPs to ascertain whether a player's performance level is "sustainable."

References

Albert, J. (2015). Improved component predictions of batting measures. arXiv preprint arXiv:1505.05557.

Bates, D., Maechler, M., Bolker, B. M., and Walker, S. (2014). lme4: Linear mixed-effects models using eigen and s4. ArXiv e-print; submitted to Journal of Statistical Software.

Brown, L. D. (2008). In-season prediction of batting averages: A field test of empirical Bayes and Bayes methodologies. The Annals of Applied Statistics, 2(1):113–152.

Carleton, R. (2012). It's a small sample size after all. Baseball Prospectus. http://www.baseballprospectus.com/article.php?articleid=17659.

FanGraphs (2015). Major league leaderboards. http://www.fangraphs.com/leaders.aspx.

FanGraphs (2016). woba. http://www.fangraphs.com/library/offense/woba/.

Friedman, J., Hastie, T. J., and Tibshirani, R. J. (2010). Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software, 33(1):1–22.

Judge, J. and BP Stats Team (2015). DRA: An in-depth discussion. Baseball Prospectus.

Neal, D., Tan, J., Hao, F., and Wu, S. S. (2010). Simply better: Using regression models to estimate major league batting averages. Journal of Quantitative Analysis in Sports, 6(3).

Null, B. (2009). Modeling baseball player ability with a nested dirichlet distribution. Journal of Quantitative Analysis in Sports, 5(2).

Tango, T. M., Lichtman, M. G., and Dolphin, A. E. (2007). The Book: Playing the percentages in baseball. Potomac Books.