True wOBA:
Estimation of true talent level for batters

Scott Powers and Eli Shayer

Stanford University

2016 SABR Analytics Conference

Which statistic is more volatile?

BABIP

$$\frac{\texttt{H - HR}}{\texttt{AB - K - HR + SF}}$$

"Stabilizes" after 820 BIP

League average: .299

OBP

$$\frac{\texttt{H + BB + HBP}}{\texttt{AB + BB + HBP + SF}}$$

"Stabilizes" after 460 PA

League average: .317

Which statistic is more volatile?

BABIP

$X_1 = \text{BABIP on } 1^{st} \text{ 150 BIP}$

$X_2 = \text{BABIP on } 2^{nd} \text{ 150 BIP}$

OBP

$Y_1 = \text{OBP on } 1^{st} \text{ 150 PA}$

$Y_2 = \text{OBP on } 2^{nd} \text{ 150 PA}$

Which is greater: $\mathbb{E}(X_1 - X_2)^2$ or $\mathbb{E}(Y_1 - Y_2)^2$?

Which statistic is more volatile?

`BABIP` `OBP`

$X_1 = \text{BABIP on } 1^{st} \text{ 150 BIP}$

$X_2 = \text{BABIP on } 2^{nd} \text{ 150 BIP}$

$Y_1 = \text{OBP on } 1^{st} \text{ 150 PA}$

$Y_2 = \text{OBP on } 2^{nd} \text{ 150 PA}$

Which is greater: $\mathbb{E}(X_1 - X_2)^2$ or $\mathbb{E}(Y_1 - Y_2)^2$?

$$\sqrt{\mathbb{E}(X_1 - X_2)^2} = 0.058$$

Which statistic is more volatile?

`BABIP` `OBP`

$X_1 =$ BABIP on 1^{st} 150 BIP

$X_2 =$ BABIP on 2^{nd} 150 BIP

$Y_1 =$ OBP on 1^{st} 150 PA

$Y_2 =$ OBP on 2^{nd} 150 PA

Which is greater: $\mathbb{E}(X_1 - X_2)^2$ or $\mathbb{E}(Y_1 - Y_2)^2$?

$\sqrt{\mathbb{E}(X_1 - X_2)^2} = 0.058$

$\sqrt{\mathbb{E}(Y_1 - Y_2)^2} = 0.062$

What is going on?

Suppose a statistic Z can be split into talent T and luck L:

$$Z_1 = T + L_1 \qquad Z_2 = T + L_2,$$

with

$$\sigma_T^2 = \mathrm{Var}(T) \qquad \text{and} \qquad \sigma_L^2 = \mathrm{Var}(L_1) = \mathrm{Var}(L_2)$$

Assuming T, L_1 and L_2 are independent,

$$\mathrm{Corr}(Z_1, Z_2) = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_L^2}$$

Outline

- Introduction
  - Regression to the mean
- Methods
  - Regularization as regression to the mean
  - True wOBA
  - Regularization vs. random effect models
- Results
  - Validation
  - Results on 2015 MLB regular season
- Discussion

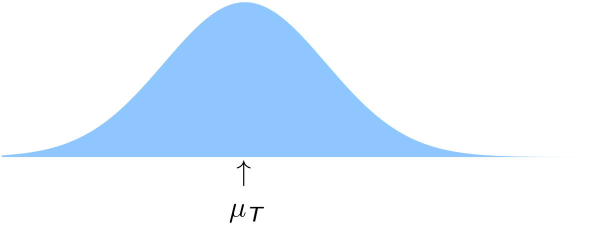
$\}$ Eli

$\}$ Scott

$\}$ Eli

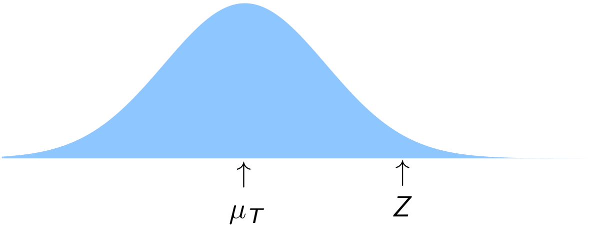
Regression to the mean

$$T \sim \mathcal{N}(\mu_T, \sigma_T^2) \qquad Z|T \sim \mathcal{N}(T, \sigma_L^2)$$



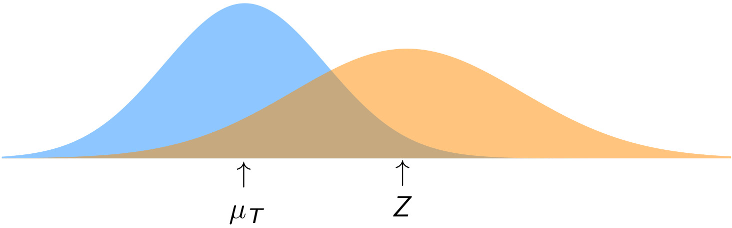
Regression to the mean

$$T \sim \mathcal{N}(\mu_T, \sigma_T^2) \qquad Z|T \sim \mathcal{N}(T, \sigma_L^2)$$



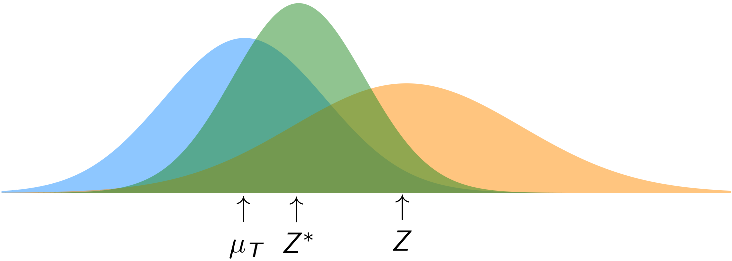
Regression to the mean

$T \sim \mathcal{N}(\mu_T, \sigma_T^2) \qquad Z|T \sim \mathcal{N}(T, \sigma_L^2)$



Regression to the mean

$$T \sim \mathcal{N}(\mu_T, \sigma_T^2) \qquad Z|T \sim \mathcal{N}(T, \sigma_L^2)$$



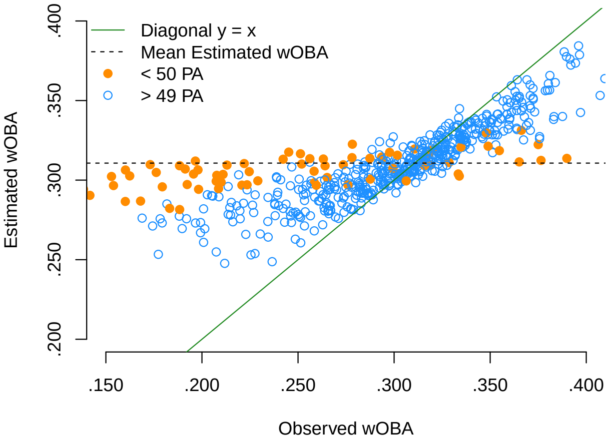
$$Z^* = \mathbb{E}[T|Z] = \frac{\sigma_T^{-2}\mu_T + \sigma_L^{-2} Z}{\sigma_T^{-2} + \sigma_L^{-2}}$$

Regression to the mean for each outcome probability

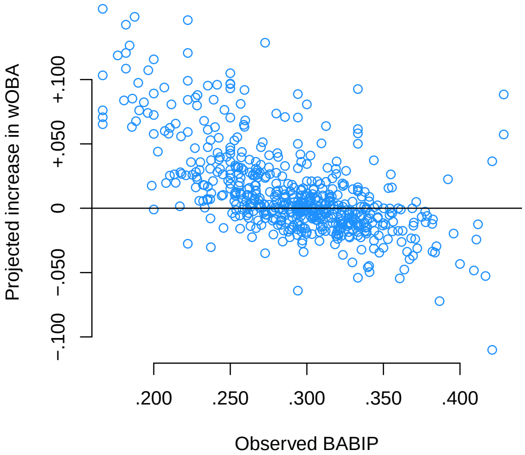
	$\hat{\sigma}_T^2$	(Naive) RMSE(Z)	(Regressed) RMSE(Z^*)
G	15.85	4.80	4.42
F	20.13	4.45	4.22
K	29.10	4.19	3.89
BB	6.26	3.33	3.04
HBP	0.24	0.94	0.80
1B	7.02	3.81	3.17
2B	0.45	2.01	1.62
3B	0.13	0.74	0.67
HR	1.88	1.79	1.61

Upshot: Different population variances for different outcomes, but regression to the mean improves RMSE for all of them!

Regressed wOBA vs. naive wOBA



Projected change in wOBA vs. BABIP



Methods

A simple linear model

Data:

For plate appearance $i \in \{1, ..., n\}$,

$$K_i = \begin{cases} 1 \text{ if } i^{th} \text{ PA results in strikeout} \\ 0 \text{ otherwise} \end{cases}$$

$B_i =$ identity of batter in i^{th} PA (e.g. Paul Goldschmidt)

Model:

$$K_i = \alpha + \beta_{B_i} + \epsilon_i, \quad \text{where} \quad \epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$$

Estimator:

$$(\hat{\alpha}, \hat{\beta}) = \arg\min \sum_{i=1}^{n} (K_i - \alpha - \beta_{B_i})^2 \quad \Rightarrow \quad \hat{\alpha} + \hat{\beta}_B = \frac{\sum_{i:B_i=B} K_i}{\sum_{i:B_i=B} 1}$$

Ridge regression

Instead of solving

$$(\hat{\alpha}, \hat{\beta}) = \arg\min \sum_{i=1}^{n} (K_i - \alpha - \beta_{B_i})^2,$$

let's try solving

$$(\alpha^*, \beta^*) = \arg\min \sum_{i=1}^{n} (K_i - \alpha - \beta_{B_i})^2 + \lambda \sum_{B} \beta_B^2, \quad \lambda > 0.$$

The result is

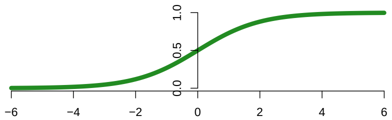
$$\beta_B^* = \frac{\lambda \cdot 0 + n_B \hat{\beta}_B}{\lambda + n_B}, \quad \text{where} \quad n_B = \sum_{i:B_i = B} 1$$

For $\lambda = \sigma_L^2 / \sigma_T^2$, this is regression to the mean!

Logistic regression

A better model:

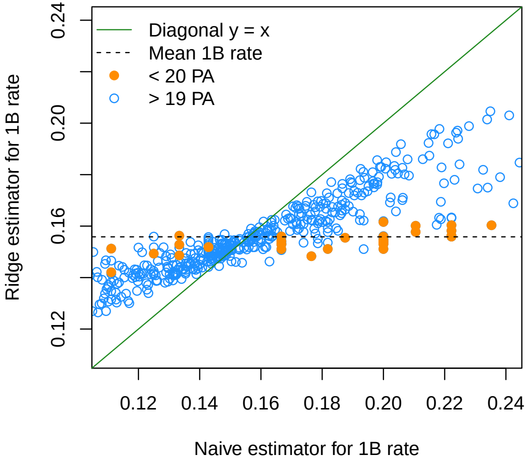
$$\eta_i = \alpha + \beta_{B_i}, \quad \text{and} \quad \mathbb{P}(K_i = 1|\eta_i) = e^{\eta_i}/(1 + e^{\eta_i})$$



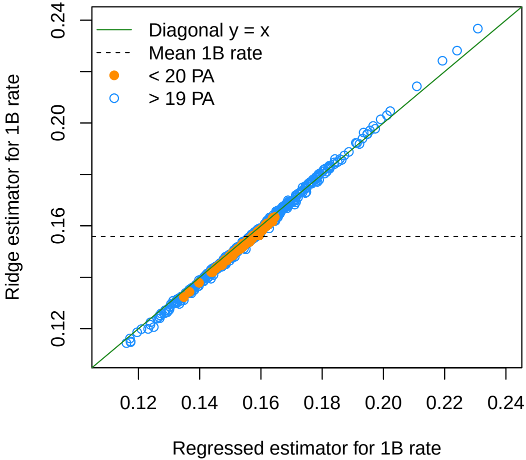
Estimator (Ridge):

$$(\alpha^*, \beta^*) = \arg\min -\sum_{i=1}^{n} \log \mathbb{P}(K_i|\eta_i) + \lambda \sum_{B} \beta_B^2$$

Ridge estimator vs. Naive estimator for 1B rate



Ridge estimator vs. Regressed estimator for 1B rate



Test RMSE: regularization vs. regression to the mean

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

Upshot: Ridge regression is essentially regression to the mean, but it allows extensions, which we will see next!

True wOBA

Data:

$Y_i \in \mathcal{Y} = \{$G, F, K, BB, HBP, 1B, 2B, 3B, HR$\}$
$B_i =$ identity of **B**atter in i^{th} PA (e.g. Paul Goldschmidt)
$P_i =$ identity of **P**itcher in i^{th} PA (e.g. Zach Greinke)
$S_i =$ identity of **S**tadium in i^{th} PA (e.g. Chase Field)
$H_i = 1$ if B_i is on **H**ome team, 0 otherwise
$O_i = 1$ if B_i and P_i have **O**pposite handedness, 0 otherwise

Model (multinomial logistic regression):

$$\eta_{ik} = \alpha_k + \beta_{B_i k} + \gamma_{P_i k} + \delta_{S_i k} + \zeta_k H_i + \theta_k O_i$$

$$\mathbb{P}(Y_i = k | \eta_i) = \frac{e^{\eta_{ik}}}{\sum_{\ell \in \mathcal{Y}} e^{\eta_{i\ell}}}$$

True wOBA

Estimation:

$$\min \left\{ -\sum_{i=1}^{n} \mathbb{P}(Y_i|\eta_i) \right.$$
$$\left. + \sum_{k \in \mathcal{Y}} \lambda_k \left(\sum_{B} \beta_{Bk}^2 + \sum_{P} \gamma_{Pk}^2 + \sum_{S} \delta_{Sk}^2 + \zeta_k^2 + \theta_k^2 \right) \right\}$$

- Choose λ_k via cross validation
- For batter B, estimated K rate in average situation is

$$\mathbb{P}_B(K) = \frac{e^{\alpha_K^* + \beta_{BK}^* + \frac{1}{2}\zeta_K^* + \frac{1}{2}\theta_K^*}}{\sum_{\ell \in \mathcal{Y}} e^{\alpha_\ell^* + \beta_{B\ell}^* + \frac{1}{2}\zeta_\ell^* + \frac{1}{2}\theta_\ell^*}}$$

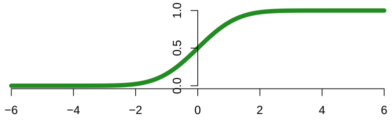
- Combine rates of outcomes into True wOBA estimate

Random effect model

Model:

$$\eta_i = \alpha + \beta_{B_i}, \quad \text{where} \quad \beta_B \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\beta^2)$$

$$\mathbb{P}(K_i = 1 | \eta_i) = \Phi(\eta_i) \leftarrow \text{Normal CDF}$$



Estimator (Random):

$$(\alpha^*, \beta^*, \sigma_\beta^{2*}) = \arg\max L(\alpha, \beta, \sigma_\beta^2 | B_i, K_i)$$

Test RMSE: regularization vs. random effect model

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

Upshot: Regularization is very similar to random effect modelling, with two differences:

- How population variance is estimated
- Regularization can be applied to multinomial regression

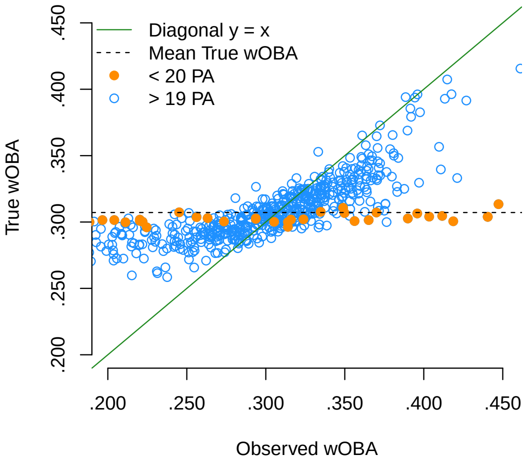
Results

Validation

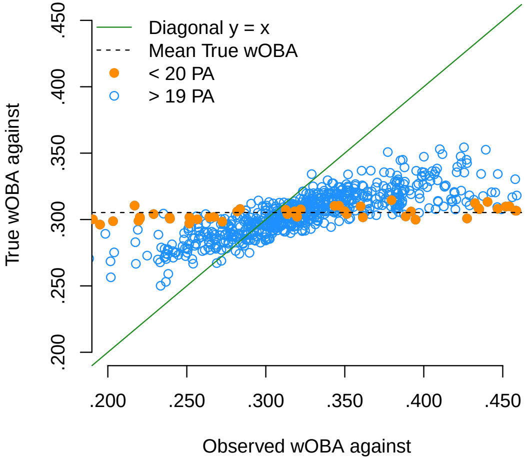
- Evaluate results on 2015 MLB regular season PAs
  - Discard intentional walks, catcher interferences
  - Discard PAs in which pitcher is batting
- Fit each method on training set to predict wOBA in test set
  - $\{O_i = 0\} \Rightarrow$ training set with prob. 90%
  - $\{O_i = 1\} \Rightarrow$ test set with prob. 90%
- Training set: 93,868 PAs
- Test set: 82,692 PAs

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

True wOBA vs. naive wOBA



True wOBA against vs. naive wOBA against



Top 5 and bottom 5 batters by True wOBA

	Batter	Team	True wOBA
	Bryce Harper	WSN	.416
Top	Mike Trout	LAA	.407
5	José Bautista	TOR	.399
	Paul Goldschmidt	ARI	.395
	Joey Votto	CIN	.393
	...		
	Alexi Amarista	SDP	.270
Bottom	Chris Owings	ARI	.269
5	René Rivera	TBR	.265
	Danny Santana	MIN	.265
	Omar Infante	KCR	.262

Top 5 and bottom 5 pitchers by True wOBA against

	Pitcher	Team	True wOBA against
Top 5	Jake Arrieta	CHC	.255
	Clayton Kershaw	LAD	.256
	Zack Greinke	LAD	.261
	Wade Davis	KCR	.267
	Dallas Keuchel	HOU	.267
	...		
Bottom 5	Jeremy Guthrie	KCR	.346
	Matt Boyd	DET	.346
	David Holmberg	CIN	.349
	Dustin McGowan	PHI	.354
	Allen Webster	ARI	.356

Top differences between naive and True wOBA

	Batter	Team	ΔwOBA
	Wilson Ramos	WSN	+.022
Top	Michael Taylor	WSN	+.021
5	Albert Pujols	LAA	+.017
	Alcides Escobar	KCR	+.016
	Chris Owings	ARI	+.014
	...		
	Anthony Rizzo	CHC	−.035
Bottom	Nolan Arenado	COL	−.037
5	Charlie Blackmon	COL	−.039
	Bryce Harper	WSN	−.045
	David Peralta	ARI	−.046

Top differences between naive and True wOBA against

	Pitcher	Team	ΔwOBA against
Top 5	Chris Rusin	COL	−.068
	Kyle Kendrick	COL	−.062
	Jerome Williams	PHI	−.047
	Matt Garza	MIL	−.045
	Kyle Lohse	MIL	−.041
	...		
Bottom 5	Jacob deGrom	NYM	+.016
	Sonny Gray	OAK	+.016
	Clayton Kershaw	LAD	+.019
	Jake Arrieta	CHC	+.021
	Zack Greinke	LAD	+.023

Discussion

Three contributions:

- We advocate use of regression to the mean instead of stabilization rates
- We explain relationship between regularized linear models and regression to the mean
- We compare regularized linear models with linear mixed effects models

Thank you!

Questions?

Scott Powers
sspowers@stanford.edu

Eli Shayer
eshayer@stanford.edu
elishayer.com

`github.com/sspowers/true-woba`

`stanfordsportsanalytics.com`