True wOBA:
Estimation of true talent level for batters

Scott Powers and Eli Shayer

Stanford University

2016 SABR Analytics Conference

Review: regression to the mean

True strikeout probability p

Observed strikeout rate $\hat{p} = \frac{\mathtt{K}}{\mathtt{PA}}$

Regression to the mean $p^* = \frac{\mathtt{K}+N\bar{p}}{\mathtt{PA}+N}$

$\bar{p} =$ league average strikeout rate

Review: regression to the mean

True strikeout probability	p	?
Observed strikeout rate	$\hat{p} = \frac{\texttt{K}}{\texttt{PA}}$	$\frac{23}{138} = 16.7\%$
Regression to the mean	$p^* = \frac{\texttt{K}+N\bar{p}}{\texttt{PA}+N}$	$\frac{23+40(20.4\%)}{138+40} = 17.5\%$

$\bar{p} =$ league average strikeout rate

Tuffy Gosewisch

Ralph Freso, Getty Images

Review: regression to the mean

True strikeout probability	p	?
Observed strikeout rate	$\hat{p} = \frac{\mathtt{K}}{\mathtt{PA}}$	$\frac{23}{138} = 16.7\%$
Regression to the mean	$p^* = \frac{\mathtt{K} + N\bar{p}}{\mathtt{PA} + N}$	$\frac{23+40(20.4\%)}{138+40} = 17.5\%$

$\bar{p}$ = league average strikeout rate

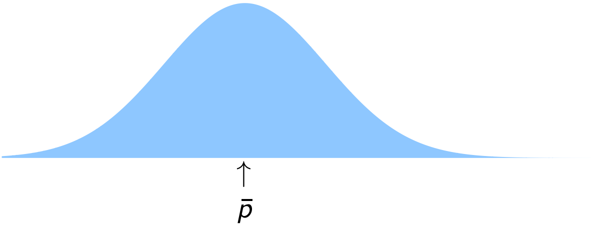
$$N = \frac{\bar{p}(1-\bar{p})}{\sigma_T^2}$$

Tuffy Gosewisch

Ralph Freso, Getty Images

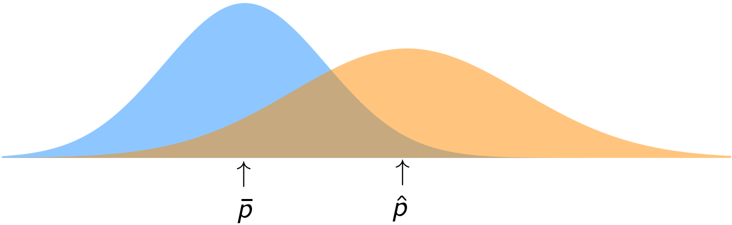
Review: regression to the mean

$p \sim \mathcal{N}(\bar{p}, \sigma_T^2)$



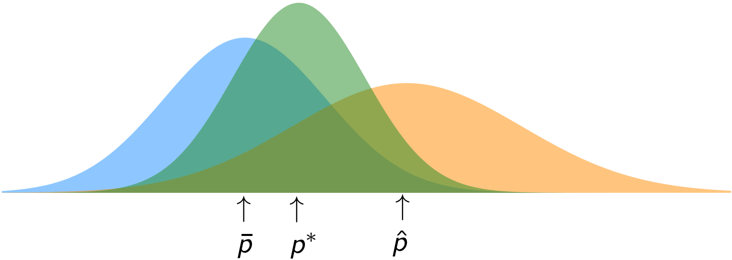
Review: regression to the mean

$$p \sim \mathcal{N}(\bar{p}, \sigma_T^2) \qquad \hat{p}|p \sim \mathcal{N}\left(p,\ \sigma_L^2 = \frac{p(1-p)}{n}\right)$$



Review: regression to the mean

$$p \sim \mathcal{N}(\bar{p}, \sigma_T^2) \qquad \hat{p}|p \sim \mathcal{N}\left(p,\ \sigma_L^2 = \frac{p(1-p)}{n}\right)$$



$$p^* = E[p|\hat{p}] = \arg\min_{p^*} E[(p-p^*)^2|\hat{p}] = \frac{\sigma_T^{-2}\bar{p} + \sigma_L^{-2}\hat{p}}{\sigma_T^{-2} + \sigma_L^{-2}}$$

Outline for this presentation

- Theory
  - ~~Regression to the mean~~
  - Regularized linear regression
  - Regularization vs. regression to the mean
  - Regularization vs. mixed effect modelling
- Application
  - Regressing wOBA to the mean
  - Comparison of true talent estimators
  - True wOBA results
- Discussion

Scott

Eli

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:

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

Regularized linear 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

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

Regularization vs. regression to the mean

Regression to the mean:

$$p^* = \frac{\sigma_T^{-2}\bar{p} + \sigma_L^{-2}\hat{p}}{\sigma_T^{-2} + \sigma_L^{-2}}$$

Regularization:

$$\alpha^* + \beta^* = \frac{\lambda\hat{\alpha} + n(\hat{\alpha} + \hat{\beta})}{\lambda + B} = \frac{\lambda\bar{p} + n\hat{p}}{\lambda + n}$$

If $\lambda = n\sigma_L^2/\sigma_T^2$, these estimates are identical!

Regularization vs. regression to the mean

Regression to the mean:

$$p^* = \frac{\sigma_T^{-2}\bar{p} + \sigma_L^{-2}\hat{p}}{\sigma_T^{-2} + \sigma_L^{-2}}$$

– σ_T^2 estimated by comparing across-player variance to σ_L^2

Regularization:

$$\alpha^* + \beta^* = \frac{\lambda\hat{\alpha} + n(\hat{\alpha} + \hat{\beta})}{\lambda + B} = \frac{\lambda\bar{p} + n\hat{p}}{\lambda + n}$$

If $\lambda = n\sigma_L^2/\sigma_T^2$, these estimates are identical!

Regularization vs. regression to the mean

Regression to the mean:

$$p^* = \frac{\sigma_T^{-2}\bar{p} + \sigma_L^{-2}\hat{p}}{\sigma_T^{-2} + \sigma_L^{-2}}$$

– σ_T^2 estimated by comparing across-player variance to σ_L^2

Regularization:

$$\alpha^* + \beta^* = \frac{\lambda\hat{\alpha} + n(\hat{\alpha} + \hat{\beta})}{\lambda + B} = \frac{\lambda\bar{p} + n\hat{p}}{\lambda + n}$$

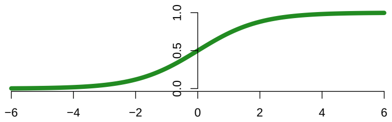
– λ chosen by cross-validation

If $\lambda = n\sigma_L^2/\sigma_T^2$, these estimates are identical!

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):

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

True wOBA

Data:

$Y_i \in \mathcal{Y} = \{\text{G, F, K, BB, HBP, 1B, 2B, 3B, HR}\}$

$B_i =$ 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:

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

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

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

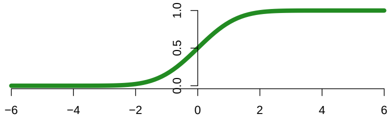
- Combine rates of outcomes into True wOBA estimate

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):

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

Application

Regression to the Mean

Regression to the mean for each outcome probability

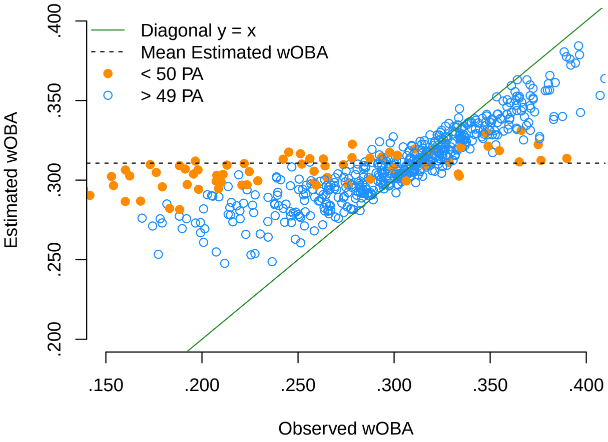
For each outcome, use 1^{st} 200 PAs to predict rate on next 200 PAs

	$\hat{\sigma}^2_T$	(Naive) RMSE($\hat{p}$)	(Regressed) RMSE(p^*)
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

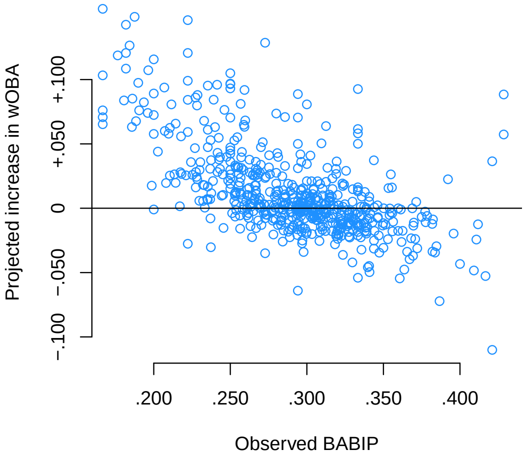
Units: percentage points

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

Regressed wOBA vs. observed wOBA



Projected change in wOBA vs. BABIP



Estimator Comparison

Test RMSE for different talent estimators

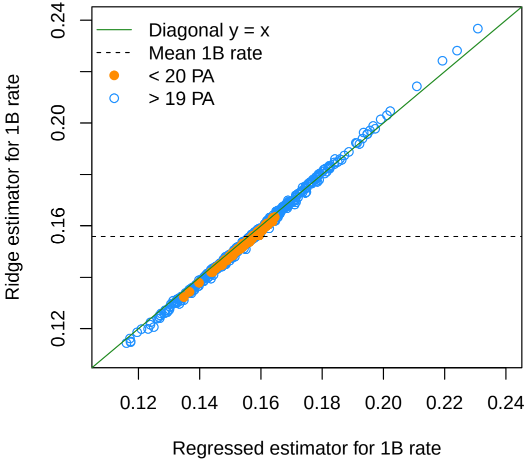
Randomly split PAs into training and test sets, using training set to predict test set rate for each outcome

	Naive	Regressed	Ridge	Random
G	4.41	3.98	3.97	3.98
F	4.45	3.97	3.99	3.98
K	4.25	3.89	3.90	3.90
BB	2.60	2.38	2.39	2.39
HBP	1.04	0.89	0.88	0.88
1B	3.66	3.09	3.08	3.08
2B	2.21	1.68	1.67	1.67
3B	0.82	0.63	0.64	0.64
HR	1.71	1.52	1.51	1.50

Units: percentage points

Upshot: In this simple example, these three estimators are virtually equivalent!

Ridge estimator vs. Regressed estimator for 1B rate



True wOBA Validation

Validation

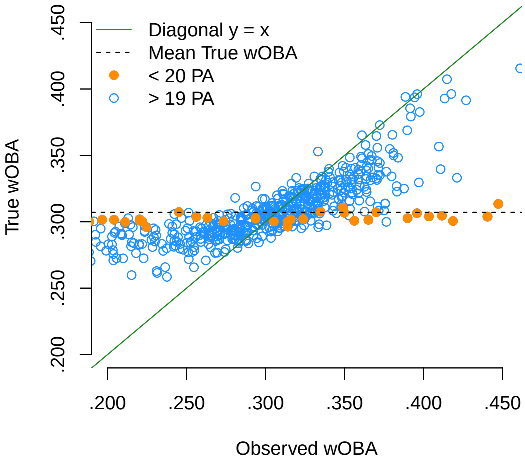
- Evaluate results on 2015 MLB regular season PAs
  - Discard intentional walks, catcher interferences
  - Discard PAs in which pitcher is batting
- Fit each method on training set to predict wOBA in test set
  - $\{O_i = 0\} \Rightarrow$ training set with prob. 90%
  - $\{O_i = 1\} \Rightarrow$ test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

Estimator	Naive	Regressed	True	Mixed
Estimated MSE	45.6	22.0	**17.3**	18.0
Stadard error	± 4.4	± 1.8	± 1.4	± 1.5

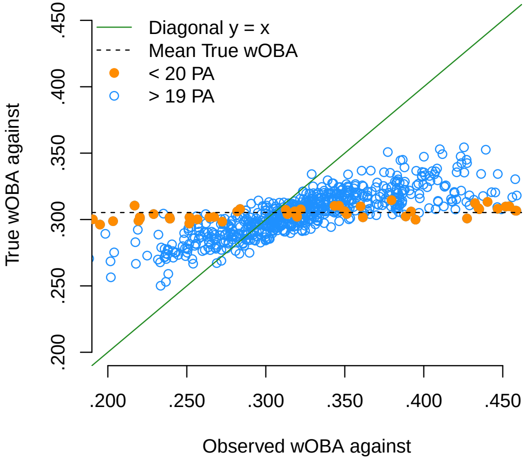
Units: wOBA points

True wOBA Results

True wOBA vs. observed wOBA



True wOBA against vs. observed wOBA against



Top 5 and bottom 5 batters by True wOBA

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

Top differences between naive and True wOBA

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

Min. 500 PA

Top differences between naive and True wOBA against

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

Min. 500 PA

Discussion

Three contributions:

- We remind everyone of regression to the mean for interpretation of small sample sizes
- 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/saberpowers/true-woba`

`stanfordsportsanalytics.com`