

Take Me Out to (Analyze) the Ballgame

Visualization and Analysis Techniques for Big Spatial Data

Chris Comiskey

May 4, 2017

Contents

1 Shiny Heat Map Confidence Intervals	1
1.1 Generalized Linear Model for Hitter Success Probabilities	1
1.2 Biomechanically Interpretable Covariates	2
1.3 Generalized Linear Model with Biomechanically Interpretable Covariates	3
1.4 Hosmer-Lemeshow Goodness of Fit Test	4
1.5 Appendix: ShinyHMCI, An R Package	4

1 Shiny Heat Map Confidence Intervals

1.1 Generalized Linear Model for Hitter Success Probabilities

As one of our research goal, we aim to create a statistical model for the heat map of success probabilities. Nonparametric methods, while straightforward, sacrifice interpretability; they achieve a modeled heat map, but without contextually interpretable components. Nonparametric models cannot relate spatially varying hitter success probabilities to hitter attributes. We propose a parametric approach using biomechanically interpretable covariates. Existing research analyzes the biomechanics of the baseball swing [Welch et al., 1995], but no research integrates those results with spatial swing outcomes in a statistical model.

Let success indicator variable, Y_{ijklm} , be a Bernoulli random variable with spatially varying mean [Sheldon et al., 2002]. Subscript $i = 1, \dots, n_{jklm}$ indexes hitter j 's swings in at bat k against pitcher l in year m .

Subscript $k = 1, \dots, n_{jlm}$ indexes hitter j 's at bats against pitcher l in year m . Subscript $l = 1, \dots, n_{jm}$ indexes pitchers hitter j faced, where n_{jm} is the total number of pitchers hitter j faced; and $m = 2007, \dots, 2016$ indexes year. Let $\mathbf{s}_{ijkl} = (px_{ikl}, pz_{ijkl}) \in \mathbf{D} \subseteq \mathbf{R}^2$ be the horizontal and vertical locations, respectively, of pitch $ijkl$ as it passes through the two dimensioned vertical face of the hitting zone. The origin, $\mathbf{s}_* = (0, 0)$, is the midpoint of the front edge of home plate, at ground level. From the pitcher's point of view, pitches to the left (right) of the center of home plate correspond to negative (positive) values of px . Pitches that bounce before reaching home plate correspond to negative values of pz .

In this study we make the simplifying assumption that location success probabilities depend on only location and hitter. This means we dispense with subscripts k, l , and m . We also assume that, given pitch location to hitter j , $\mathbf{s}_{ij} = (px_{ij}, pz_{ij})$, swings are independent Bernoulli trials. This gives $Y_{ij} | \mathbf{s}_{ij} \sim \text{Bernoulli}(p_{ij})$, where $E[Y_i | \mathbf{s}_{ij}] = p_{ij}$.

Accordingly, let $i = 1, \dots, n_j$ index hitter j 's swings, out of n_j total swings on record. Let $\mathbf{X}_{ij}(\mathbf{s}_{ij})$ be covariates specific to hitter j and location \mathbf{s}_{ij} on swing i . A Bernoulli random variable suggests a generalized linear model with logit link function for relating success probability to covariate information:

$$\text{logit}(p_{ij} | \mathbf{X}_{ij}(\mathbf{s}_{ij})) = \mathbf{X}_{ij}(\mathbf{s}_{ij})\boldsymbol{\beta}_j, \quad (1)$$

where $\boldsymbol{\beta}_j$ is the vector of covariate coefficient parameters specific to hitter j [Myers et al., 2012]. Next, we discuss and develop covariates.

1.2 Biomechanically Interpretable Covariates

Why does Peralta, and why do hitters in general, hit pitches in some locations better than others? We submit biomechanics as potentially part of the answer. Biomechanics underpin why hitters prefer some pitch locations more than others. Given the choice, athletes select a specific place for the ball before swinging. Consider golf, a sport where the ball is stationary, and the acting athlete chooses where to stand in relation to the ball. In fact, golfers position themselves very precisely in relation to the ball to achieve impact at the optimal point in their swing [Cochran and Stobbs, 2005]. If the impact point deviates from the ideal location, performance suffers. Consider tennis, a step closer to baseball, in that the ball approaches, but the player has time to position himself relative to the incoming ball. Once again, tennis players strive to hit the ball at a specific point in their forehand, a precise distance from the ground and from their body [Elliott, 2006]. As with golf, if the point of impact deviates from this location, performance suffers. Note

that in both sports the ideal player to ball positioning depends on, at the very least, anatomy, biomechanics, and equipment. We submit the same dynamics affect baseball hitting. However, in baseball the hitter cannot predetermine ball location, nor does he have time to reposition himself in response to the location and trajectory of the incoming pitch. For these reasons, meaningful measurements of hitter to ball distance and angle are reasonable covariates. Polar coordinate pitch locations would inherently provide this type of meaningful covariate for use and interpretation in our models.

To illustrate, in Figure 3 we shift the origin to a hitter's approximate center of gravity in his stance, where the extended bat line intersects his axis of rotation at the moment of contact [Welch et al., 1995].

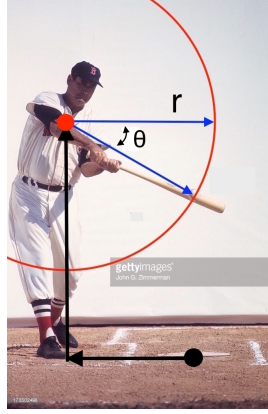


Figure 1: The ground level black dot represents the origin (0,0) in the rectangular coordinate system. The translated origin (red dot) coincides with the hitter's approximate center of gravity, and thus the polar origin. The bold arrows show the origin shift. The length of the arrows moving out from his center of gravity to specific grid locations represent r , and the angle between the same two arrows represent θ .

Referring to Figure 3, let r measure the distance from the hitter's center of gravity to the ball at impact, and let θ be the angle below horizontal of the line segment connecting the center of gravity and the ball at impact. As in golf and tennis, ball location—too far/close to the hitter, or above/below the ideal point of impact— affects hitting performance. Letting $\mathbf{X}_{ij}(\mathbf{s}_{ij})$ in (1) be comprised of r_{ij} and θ_{ij} terms provides an exploratory starting point.

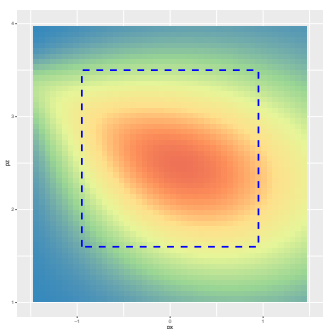
1.3 Generalized Linear Model with Biomechanically Interpretable Covariates

Let covariate vector $\mathbf{X}_{ij}(\mathbf{s}_{ij})$ in (1) be defined as $\mathbf{X}_{ij}(\mathbf{s}_{ij}) = \{r_{ij}, \theta_{ij}, r_{ij}\theta_{ij}, r_{ij}^2, \theta_{ij}^2, r_{ij}^2\theta_{ij}^2\}$. Substituting into (1) yields:

$$\text{logit}(p_{ij}|\mathbf{s}_{ij}, r_{ij}, \theta_{ij}) = \beta_{j0} + \beta_{j1}r_{ij} + \beta_{j2}\theta_{ij} + \beta_{j3}r_{ij}\theta_{ij} + \beta_{j4}r_{ij}^2 + \beta_{j5}\theta_{ij}^2 + \beta_{j6}r_{ij}^2\theta_{ij}^2 \quad (2)$$

Note that given a hitter j , and pitch location \mathbf{s}_{ij} , the elements of \mathbf{X}_{ij} are simply a trigonometric function of \mathbf{s}_{ij} and the translated origin. Thus, for convenience, we replace $\text{logit}(p_{ij}|\mathbf{s}_{ij}, r_{ij}, \theta_{ij})$ with $\text{logit}(p_{ij}|\mathbf{s}_{ij})$ for the remainder of this study.

We choose Johnny Peralta from Chapter 1 to illustrate, and let $j = P$ for convenience. We fit model (2) using Peralta's $n_P = 9177$ observed swings, find maximum likelihood estimates of β_P using an iteratively reweighted least squares algorithm [Myers et al., 2012].



Covariate	β_i	MLE	SE	p
N/A	β_0	-4.08	0.70	< 0.001
r	β_1	1.19	0.51	0.018
θ	β_2	-1.93	1.90	0.311
$r * \theta$	β_3	-1.64	0.70	0.064
r^2	β_4	-0.32	0.09	< 0.001
θ^2	β_5	-3.92	1.10	< 0.001
$r^2 * \theta^2$	β_6	-0.46	0.21	0.025

Figure 2 & Table 1: A table beside a figure

1.4 Hosmer-Lemeshow Goodness of Fit Test

Generalized Linear Models [Myers et al., 2012]

- (pg 147) Hosmer-Lemeshow test = Logistic regression Goodness of Fit test.
- Like Pearson Chi-Sq, but for continuous covariates.
- Order all responses according to fitted \hat{p} , then group into deciles.
- Then basically $\chi^2 = \sum \sum \frac{(O-E)^2}{E}$
- p-value = 0.1513

1.5 Appendix: ShinyHMCI, An R Package

References

Christian M Welch, Scott A Banks, Frank F Cook, and Pete Draovitch. Hitting a baseball: A biomechanical description. Journal of Orthopaedic & Sports Physical Therapy, 22(5):193–201, 1995.

Ross Sheldon et al. A first course in probability. Pearson Education India, 2002.

Raymond H Myers, Douglas C Montgomery, G Geoffrey Vining, and Timothy J Robinson. Generalized linear models: with applications in engineering and the sciences, volume 791. John Wiley & Sons, 2012.

Alastair J Cochran and John Stobbs. Search for the perfect swing. Triumph, 2005.

B Elliott. Biomechanics and tennis. British Journal of Sports Medicine, 40(5):392–396, 2006.