Swinging, Fast and Slow: Untangling intention and timing error from bat speed and swing length in Major League Baseball

Scott Powers[1] and Ronald Yurko[2]

[1]Department of Sport Management, Rice University
[2]Department of Statistics & Data Science, Carnegie Mellon University

Introduction

In May 2024, Major League Baseball (MLB) released a novel dataset of pitch-level swing tracking metrics. The data are limited to bat speed and swing length at contact point, with a teaser of more to come.

Interpretation of bat speed and swing length is complicated by the measurement point. For the exact same swing mechanics, if the batter swings later (or the ball moves faster), the contact point occurs upstream in the swing path. This results in a shorter swing length measurement and a slower bat speed measurement (because the bat accelerates up to and beyond intended contact). If the batter swings earlier (or the ball moves slower), the contact point occurs downstream in the swing path, causing a different measurement bias. Bat speed and swing length, therefore, reflect not only the swing but also the outcome of the swing.

This complication has not discouraged data scientists from making claims about what constitutes a *good* swing and how batters *should* try to swing. In the present work, we take a more cautious approach, seeking to understand and classify the sources of swing-to-swing variability in bat speed and swing length.

Data

Our dataset comprises 96,887 swings by MLB players between April 3 and May 30, 2024. For each swing, we observe bat speed and swing length. **Bat speed** is the linear speed of the *sweet spot* of the bat (approximately six inches from the end), measured at the point of contact with the ball. If no contact occurs, the contact point is defined as the point of minimum distance between ball and sweet spot. **Swing length** is the distance travelled by the end of the bat from start of swing to contact point. In addition to these novel swing tracking metrics, we have ball tracking data (most notably pitch speed, pitch location and batted ball speed) and contextual data (most notably ball-strike count). All data are publicly available from Baseball Savant.

Methods

Cleaning the data. The initial dataset includes attempted bunts (which are labelled) and attempted *check-swings* (i.e., partial swings, not labelled). We aim to limit our analysis to full swings. When reporting aggregate metrics like average bat speed, MLB filters out each batter's bottom 10% of swings by bat speed, labelling the remaining swings as *competitive*. We take a more inclusive approach by filtering out only swings with bat speed below 50 mph (which filters out all bunt attempts). Based on video review, this cutoff does a good job of distinguishing between full and partial swings while only removing $\approx 2.5\%$ of swings.

We reproduce MLB's derived metric for *squared-up* contact as:

$$\text{batted ball speed} > 80\% \times \{1.23 \cdot (\text{bat speed}) + 0.23 \cdot (\text{pitch speed})\}\,. \tag{1}$$

The right-hand side of (1) is a rough approximation for 80% of the theoretical maximum batted ball speed given bat speed and pitch speed at contact (with additional assumptions about the mass of bat and ball). In our dataset, we observe batted ball speeds as high as 110% of the rough theoretical maximum.

Predicting swing intention. We begin with the hypothesis that two covariates explain real (non-artifactual) differences in swing mechanics: ball-strike count and pitch location. To the extent that a batter's swing tracking metrics co-vary with count, we describe this as their *approach*. To the extent that a batter's swing tracking metrics co-vary with pitch location, we describe this as swing *adaptation*. Acknowledging that swing timing biases the measurement of bat speed and swing length, we mitigate this confounding bias by **filtering on swings that result in squared-up contact** and fit a linear mixed-effects model:

$$\begin{aligned}(\text{swing length})_i = \alpha + \gamma^A_{b_i} + (\beta^B + \gamma^B_{b_i}) \cdot (\text{balls})_i + (\beta^S + \gamma^S_{b_i}) \cdot (\text{strikes})_i \\ + (\beta^X + \gamma^X_{b_i}) \cdot (\text{pitch loc x})_i + (\beta^Z + \gamma^Z_{b_i}) \cdot (\text{pitch loc z})_i + \epsilon_i, \qquad (2)\end{aligned}$$

$$\gamma^A_b \sim \mathcal{N}(0, \sigma^2_A), \;\; \gamma^B_b \sim \mathcal{N}(0, \sigma^2_B), \;\; \gamma^S_b \sim \mathcal{N}(0, \sigma^2_S), \;\; \gamma^X_b \sim \mathcal{N}(0, \sigma^2_X), \;\; \gamma^Z_b \sim \mathcal{N}(0, \sigma^2_Z), \;\; \epsilon_i \sim \mathcal{N}(0, \sigma^2_\epsilon),$$

where b_i is the batter swinging at pitch i. The γ parameters are random intercepts and random slopes. We fit a model with the same specification to predict intended bat speed. These models tell us: What are the bat speed and swing length (by batter, count and pitch location) when the timing is good? We interpret the prediction from this model on each pitch as the *intended* bat speed and swing length.

Estimating swing timing. We define timing as the standard deviation of residual (observed minus intended) bat speed and swing length. This is a two-dimensional measurement with units of miles per hour and feet. We hypothesize that smaller values in both dimensions (i.e., less timing error) are better.

Results

Figure 1 demonstrates the practical significance of separating the variation due to intention from the variation due to timing error. Between intended bat speed and swing length, we see an approximately linear relationship. However, we see a very different relationship between residual (observed minus intended) bat speed and swing length. This curve seems to trace the speed of the bat along the path of the swing, accelerating until shortly after the intended contact point and then slowing down. In addition, we have interpretable metrics describing each batter's approach, swing adaptation and timing (not shown in this abstract).

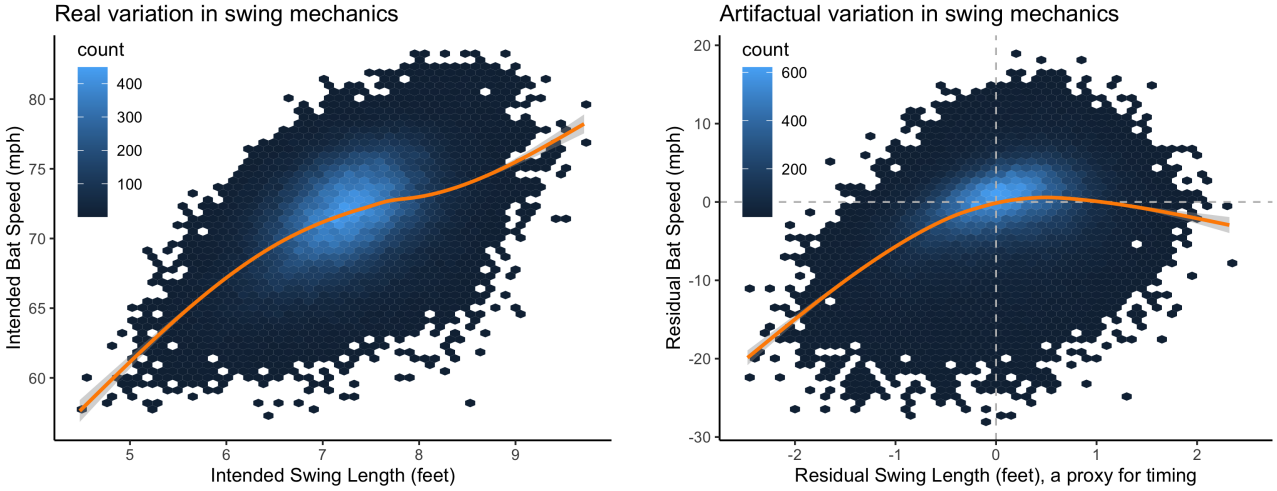

Figure 1. (left) Intended bat speed vs. swing length. (right) Residual (observed minus intended) bat speed vs. swing length.

Conclusion

By predicting batter intention, we mitigate the effect of timing error on measuring bat speed and swing length. It is too early to conclude that our timing metric is a valid measurement of swing timing. As Figure 1 shows, there is still much variation in residual bat speed that remains unexplained by residual swing length. These improved measurements of real (non-artifactual) variation in swing mechanics will enable future research on topics such as the effect of approach on results.