A Run Expectancy Approach to Lead Distance Optimization in Major League Baseball

Jack Whitney-Epstein[1], Zach Sissman[2], and Lila Dodson[3]

[1]The Brunswick School, Greenwich, CT
[2]Community School of Naples, Naples, FL
[3]San Francisco University High School, San Francisco, CA

October 31, 2025

Abstract

A runner's primary lead off first base creates leverage to steal second but also exposure to pickoffs. We develop a nested sequence of logistic models to estimate (i) pickoff attempts, (ii) pickoff success given an attempt, (iii) steal attempts given no pickoff, and (iv) steal success given an attempt, using 2024 MLB data and Baseball Savant metrics. We map stage probabilities to expected runs via fixed linear weights (+0.20 for a successful steal; -0.45 for caught stealing or picked off) and optimize over lead distance to obtain a context-specific optimal lead L^*. Empirically, observed leads are modestly larger than optimal on average (+0.19 ft), with a larger gap on steal attempts (+0.67), consistent with unobserved intent to steal. This framework quantifies the central trade-off – greater leads increase steal success but raise pickoff risk – on a common expected-runs scale and yield actionable, interpretable recommendations within the observed support.

1 Introduction

1.1 History of Base Stealing

In baseball, scouts traditionally evaluate five tools – hitting for average, hitting for power, fielding, throwing, and speed. Among these, *speed* is often underappreciated, yet it strongly shapes a player's impact on the base paths. Stolen bases are the most conspicuous expression of that impact.

Historically, steals have waxed and waned with the run-scoring environment of the MLB. During the Deadball era (c. 1900–1920), home run rates were quite low and teams relied more on advancing runners to score. Many seasons saw clubs exceed 100 steals and leagues tally well over 1,000 steals combined (McMurray, 2015). As home runs rose from the 1930s through the 1950s, average team steals fell to roughly 39 per season (Baseball-Reference, nd). Then from the 1960s to the 1980s, players like Lou Brock and Rickey Henderson brought

base-stealing back into fashion, and Maury Wills's remarkable 1962 season (featuring 104 stolen bases) was a turning point in baserunning that influenced how opponents pitched and held runners on (Vazzana, 2016).

1.2 Sabermetrics and Expected Runs

By the 1990s–2000s, **sabermetrics** reframed the value of the steal. Bill James argued that a steal must succeed about two-thirds of the time to break even – a "rule of thumb" that recognizes outs as scarce resources and emphasizes efficiency over raw totals (James, 2023). Building on this, *linear weights* assign average run values to outcomes independent of specific context. On Baseball Savant, a successful steal is valued at +0.20 expected runs (or xRuns), while being caught stealing or picked off are -0.45 xRuns each (Baseball Savant, nd). These provide a consistent baseline for evaluation.

However, context still matters when deciding to steal a base. The *run value* of a steal depends on the *base-out state* (which bases are occupied × number of outs). Note that there are $2^3 \times 3 = 24$ possible such states. Sabermetrics defines the "value of a play" as the *change in run expectancy* between states, and run expectancy tables (like Table 1) are frequently used to estimate this. According to Table 1, stealing second base with 0 outs and no one else on base is worth $1.068 - 0.831 = 0.237$ runs, while being caught stealing in the same situation is worth $0.243 - 0.831 = -0.588$ runs (FanGraphs, nd).

Table 1: Example run expectancy table

Runners	0 Outs	1 Out	2 Outs
Empty	0.461	0.243	0.095
1 _ _	0.831	0.489	0.214
_ 2 _	1.068	0.644	0.305
1 2 _	1.373	0.908	0.343
_ _ 3	1.426	0.865	0.413
1 _ 3	1.798	1.140	0.471
_ 2 3	1.920	1.352	0.570
1 2 3	2.282	1.520	0.736

Example run expectancy table (FanGraphs, nd).

1.3 The Lead's Impact on Base Stealing

While speed is a key ingredient in stealing bases, the *lead* a runner takes off first base critically shapes both *steal success* and *pickoff risk*. A larger lead shortens the distance to second if the runner goes, but it also invites (and increases the success of) pickoff attempts at first.

This strategic dimension of stealing has been studied in game-theoretic terms. Turocy (2014) models the steal decision as a two-player inspection game between the runner and the defense, where the runner chooses whether to go and the defense allocates attention between the

runner and the batter. The mixed-strategy equilibrium implies a near-linear relationship between attempt frequency and success rate, with validation over MLB play-by-play data from 1974–2011. However, the framework does not prescribe an *optimal primary lead* nor incorporate pitcher- and catcher-specific skills at the micro level.

A complementary physics approach examines the kinematics of a steal attempt. Using position–time and velocity–time profiles for a case study, Kagan (2013) parametrizes acceleration from first, slide deceleration, top speed, speed upon arrival, and the runner's lead distance. He finds that sprint speed and initial acceleration most strongly affect success, with comparatively modest effects attributed to the primary lead. This analysis, while insightful mechanically, only implicitly considers pickoff risk and is limited in its scope.

Data-driven work using tracking-era measurements shows how player behavior and pitcher tendencies shape leads. Lindbergh (2015) uses MLB data on primary/secondary leads and sprint speed, highlighting outliers (e.g., Jon Lester's reluctance to throw over, enabling larger leads; Ichiro Suzuki's larger-than-expected leads relative to speed) and noting a positive association between sprint speed and lead size. That study, however, does not jointly model runner, pitcher, and catcher pop time, nor does it connect the full decision sequence (pickoff attempt $\rightarrow$ pickoff success; steal attempt $\rightarrow$ steal success) to *expected runs*.

This study extends existing literature by (i) modeling the baserunning decision sequence with a nested set of logistic regressions, (ii) incorporating runner sprint speed, pitcher hold ability (*Threat*), and catcher pop time in the relevant stages, and (iii) optimizing *primary lead distance* to maximize expected runs for the base state of a runner on first (second and third empty). In doing so, we jointly quantify the reward of a larger lead (higher steal probability) and its cost (higher pickoff probability) on a common expected-runs scale.

2 Methodology

2.1 Data Overview

Our primary dataset, provided by MLB, comprises 2024 play-by-play data. It includes all pickoff attempts, a random sample of called pitches, and every stolen base attempt on takes (pitches where the batter did not swing). Each observation includes pitch context (inning, date, home/away, count, outs), participant IDs (runner, pitcher, catcher), and the runner's primary and secondary lead distances in feet (we denote the primary lead by L). To avoid base-state confounding, we restrict to the state with a runner on first and second/third empty. We then merge in the following covariates from Baseball Savant that may influence the decision to steal, namely:

- Runner sprint speed, s (in feet/second).
- Catcher pop time, p (seconds): the average time it takes a catcher to throw the ball to second base after receiving a pitch
- Pitcher "Threat," θ (per 100 IP): Baseball Savant's Net Bases Prevented (Baseball Savant, nd) scaled to 100 innings pitched, $\theta = 100\,\frac{\text{NBP}}{\text{IP}}$; higher values indicate better

control of the running game.

To evaluate the expected value of base-stealing outcomes, we used *fixed linear weights* from Baseball Savant:

$$\text{xRuns} = 0.20 \cdot \mathbf{1}\{\text{SB}\} - 0.45 \cdot \mathbf{1}\{\text{CS}\} - 0.45 \cdot \mathbf{1}\{\text{PK}\} \tag{1}$$

where $\mathbf{1}\{A\}$ represents an indicator that is 1 if A occurs and 0 otherwise. This metric estimates the relative value of a stolen base (SB) and the harmful cost of being picked off (PK) or caught stealing (CS).

2.2 Modeling Framework

We proceed by representing the baserunning process as a *nested sequence of conditional events* once a runner reaches first, namely:

- Pickoff attempt (PO).
- If attempted: pickoff successful (PK | PO).
- If no pickoff: steal attempt (ATT | ¬PO).
- If attempted: successful steal (SB | ATT).

If none of these occur, the pitch is classified as **Nothing**. Each stage is modeled with a *logistic regression* with predictors reflecting the hypothesized mechanisms for each stage. Formally,

$$\mathbb{P}(\text{PO} \mid L, \theta) = \text{logit}^{-1}(\alpha_0 + \alpha_1 L + \alpha_2 \theta), \tag{2}$$

$$\mathbb{P}(\text{PK} \mid \text{PO}, L, \theta) = \text{logit}^{-1}(\beta_0 + \beta_1 L + \beta_2 \theta), \tag{3}$$

$$\mathbb{P}(\text{ATT} \mid \neg\text{PO}, \theta, p, s) = \text{logit}^{-1}(\gamma_0 + \gamma_1 \theta + \gamma_2 p + \gamma_3 s), \tag{4}$$

$$\mathbb{P}(\text{SB} \mid \text{ATT}, L, \theta, p, s) = \text{logit}^{-1}(\delta_0 + \delta_1 L + \delta_2 \theta + \delta_3 p + \delta_4 s). \tag{5}$$

Note that in Equation 4, we intentionally exclude the primary lead L and treat $\mathbb{P}(\text{ATT} \mid \neg\text{PO}, \theta, p, s)$ as the runner's *baseline green-light probability of stealing 2nd base*. This avoids simultaneity of L and (unobserved) intent to steal, implying that L influences attempt probability only indirectly through its effect on $\mathbb{P}(\text{PO} \mid L, \theta)$.

Now letting $X = (\theta, p, s)$ represent known player characteristics, conditional probability laws give us the following:

$$\mathbb{P}(\text{PK} \mid L, \theta) = \mathbb{P}(\text{PK} \mid \text{PO}, L, \theta) \cdot \mathbb{P}(\text{PO} \mid L, \theta)$$

$$\mathbb{P}(\text{ATT} \mid L, X) = \mathbb{P}(\text{ATT} \mid \neg\text{PO}, \theta, p, s) \cdot (1 - \mathbb{P}(\text{PO} \mid L, \theta))$$

$$\mathbb{P}(\text{SB} \mid L, X) = \mathbb{P}(\text{SB} \mid \text{ATT}, L, X) \cdot \mathbb{P}(\text{ATT} \mid L, X)$$

$$\mathbb{P}(\text{CS} \mid L, X) = (1 - \mathbb{P}(\text{SB} \mid \text{ATT}, L, X)) \cdot \mathbb{P}(\text{ATT} \mid L, X)$$

Taking conditional expectations of Equation 1 gives the following:

$$\mathbb{E}[\text{xRuns} \mid L, X] = 0.20 \cdot \mathbb{E}[\mathbf{1}\{\text{SB}\} \mid L, X] - 0.45 \cdot \mathbb{E}[\mathbf{1}\{\text{CS}\} \mid L, X] - 0.45 \cdot \mathbb{E}[\mathbf{1}\{\text{PK}\} \mid L, X],$$

which implies that

$$\text{xRuns}(L, X) = 0.20 \cdot \mathbb{P}(\text{SB} \mid L, X) - 0.45 \cdot \mathbb{P}(\text{CS} \mid L, X) - 0.45 \cdot \mathbb{P}(\text{PK} \mid L, X), \tag{6}$$

which is fully estimable from our logistic models in Equations 2 to 5.

2.3 Objective and Optimization

Then for a fixed triplet $X = (\theta, p, s)$ corresponding to a specific pitcher-catcher-runner situation, we define expected runs as a function of the lead:

$$\text{xRuns}(L) = 0.20 \cdot \mathbb{P}(\text{SB} \mid L) - 0.45 \cdot \mathbb{P}(\text{CS} \mid L) - 0.45 \cdot \mathbb{P}(\text{PK} \mid L) \tag{7}$$

We then compute the *optimal lead* for that situation as

$$L^* = \arg\max_{L \in \mathcal{L}} \text{xRuns}(L), \tag{8}$$

using `R`'s `optimize()` function over a bounded domain $\mathcal{L}$ restricted to the *observed support of leads in our data*. We then compare the observed lead to L^* to quantify *lead deviation*.

3 Results

3.1 Model Estimates

We report coefficient estimates for our logistic regressions in Tables 2–5. All interpretations below are *associative (not causal)*, conditional on the included covariates and model specification.

In the pickoff-attempt model (Table 2), a one-foot larger lead is *associated with* roughly 20% higher odds of a throw-over ($\exp(0.1863) \approx 1.20$), while a one-unit increase in pitcher *Threat* is associated with about 1.2% lower odds ($\exp(-0.0125) \approx 0.988$), holding lead fixed.

In the pickoff-success-given-attempt model (Table 3), pickoff success is positively associated with both lead distance (79% higher odds per foot; $\exp(0.5844) \approx 1.79$) and *Threat* (19% higher odds per unit; $\exp(0.1763) \approx 1.19$).

In the steal-attempt-given-no-pickoff model (Table 4), a one-unit increase in *Threat* is associated with a 9% decrease in attempt odds ($\exp(-0.0969) \approx 0.91$). Attempt odds are higher with slower catchers (about +12% per 0.1 s of pop time; $\exp(0.11298) \approx 1.12$) and faster runners (about +85% per +1 ft/s of sprint speed; $\exp(0.6153) \approx 1.85$).

In the steal-success-given-attempt model (Table 5), success odds are higher with longer leads (about +8% per foot; $\exp(0.0769) \approx 1.08$) and slower pop times (about +68% per 0.1 s;

$\exp(0.5213) \approx 1.68$), lower with higher *Threat* (about –6% per unit; $\exp(-0.0603) \approx 0.94$), and only weakly related to sprint speed ($p \approx 0.08$) once other factors are included.

Overall, these associations align with baseball intuition and indicate that the models recover expected trade-offs between lead size, pickoff risk, and steal success. Note that these estimates are associative – not causal – and should be interpreted conditional on the included covariates and model specification.

Table 2: Logistic regression for pickoff attempts (PO)

Term	Estimate	Std. Error	z value	$\Pr(>\|z\|)$
(Intercept)	-4.1343	0.0737	-56.072	$< 2 \times 10^{-16}$
PrimaryLead1B	0.1863	0.0068	27.501	$< 2 \times 10^{-16}$
Threat	-0.0125	0.0027	-4.647	3.37×10^{-6}

Table 3: Logistic regression for pickoff success given attempt (PK | PO)

Term	Estimate	Std. Error	z value	$\Pr(>\|z\|)$
(Intercept)	-10.9581	0.6992	-15.672	$< 2 \times 10^{-16}$
PrimaryLead1B	0.5844	0.0573	10.200	$< 2 \times 10^{-16}$
Threat	0.1763	0.0263	6.698	2.11×10^{-11}

Table 4: Logistic regression for steal attempt given no pickoff (ATT | ¬PO)

Term	Estimate	Std. Error	z value	$\Pr(>\|z\|)$
(Intercept)	-22.6976	0.9395	-24.159	$< 2 \times 10^{-16}$
Threat	-0.0969	0.0036	-27.223	$< 2 \times 10^{-16}$
poptime	1.1298	0.3976	2.841	4.49×10^{-3}
sprint_speed	0.6153	0.0186	33.118	$< 2 \times 10^{-16}$

3.2 Expected Runs and Lead Deviation

For each pitcher-catcher-runner triplet $X_i = (\theta, p, s)$ and each candidate lead distance $L \in \mathcal{L}$, we compute $\text{xRuns}(L \mid X_i)$ from our fitted models (Section 3.1) via Equation 6. We then report the *optimal lead* as in Equation 8:

$$L^* = \arg\max_{L \in \mathcal{L}} \text{xRuns}(L \mid X_i)$$

We define *lead deviation* as

$$d_i \equiv L_i^{\text{obs}} - L_i^*, \tag{9}$$

so $d_i > 0$ indicates an aggressive (too-large) lead and $d_i < 0$ a conservative (too-small) lead. We compute this deviation for all observations in our dataset and display the distribution

Table 5: Logistic regression for steal success given steal attempt (SB | ATT)

Term	Estimate	Std. Error	z value	$\Pr(>\|z\|)$
(Intercept)	-12.2005	2.2193	-5.498	3.85×10^{-8}
PrimaryLead1B	0.0769	0.0351	2.193	2.83×10^{-2}
Threat	-0.0603	0.0108	-5.594	2.22×10^{-8}
poptime	5.2134	0.9071	5.747	9.08×10^{-9}
sprint_speed	0.0770	0.0442	1.743	8.14×10^{-2}

of lead deviations in Figure 1. This distribution is approximately normal and is centered at +0.19 ft. Additionally, we plot the distribution of lead deviations when a stolen base is attempted in Figure 2. This distribution is also approximately normal, but with a higher mean deviation from our estimated optimal lead (+0.67 ft). This is very likely indicative of *intent to steal*, which is unknown to us. Additional visualizations of lead deviation are included in Appendix A.

3.3 Case Study: Cubs vs. Pirates, August 28, 2024

In an August 28, 2024 game between the Chicago Cubs and Pittsburgh Pirates, Pete Crow-Armstrong (PCA) took an 11.46 foot lead against opposing pitcher Paul Skenes and catcher Yasmani Grandal (pictured in Figure 3). Considering PCA's sprint speed of 30 ft/s, Skenes' estimated Threat of 4.039, and Grandal's pop time of 2.09 seconds, our model produces an optimal lead estimate (L^*) of 10.37 ft. PCA's lead exceeds this estimate by 1.09 feet, indicating a slightly more aggressive lead than recommended.

Figure 4 displays the model-implied probabilities of SB, CS, and PK as functions of primary lead L given the players involved, as well as the expected-runs estimate xRuns(L). Figure 5 demonstrates that a lead distance of 10.37 ft maximizes estimated expected runs (+0.007). This turning point reflects the risk-reward balance when taking a lead: larger leads increase the probability of reaching second safely, at the cost of increased pickoff danger.

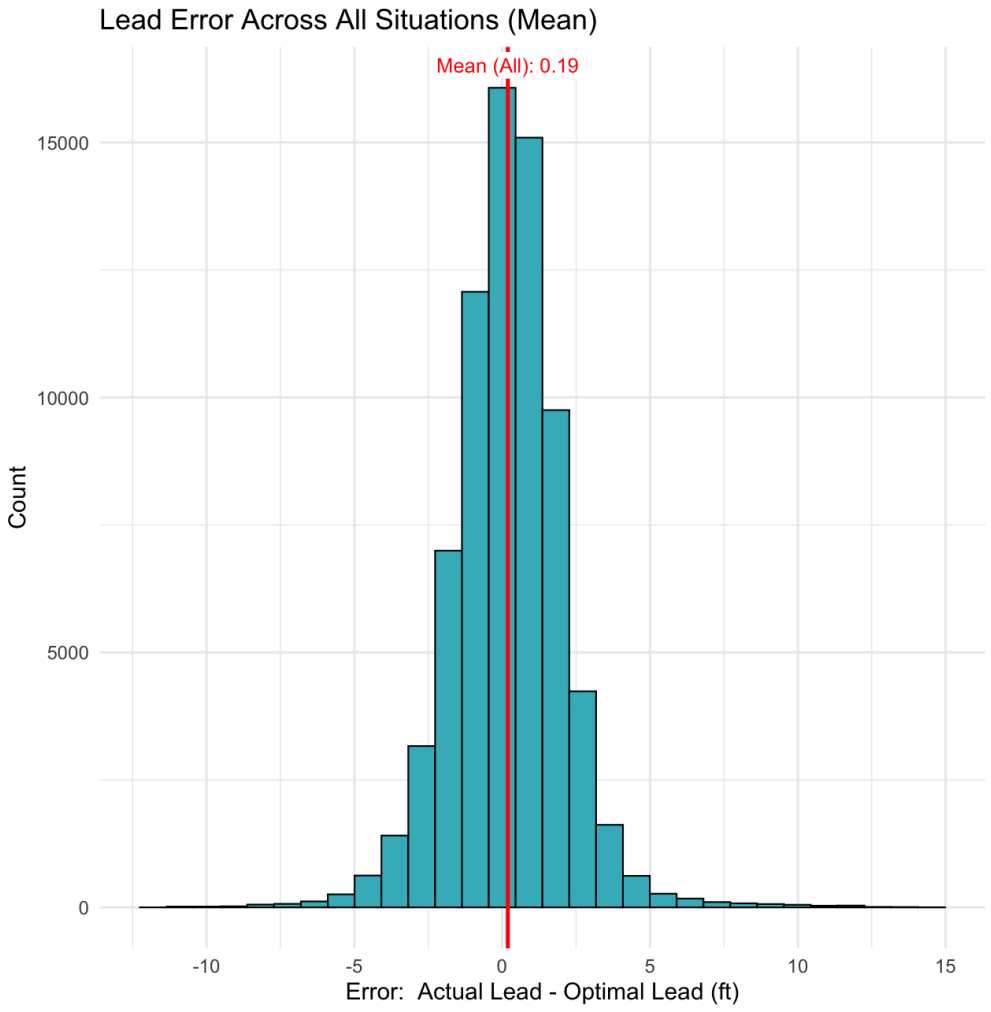


Figure 1: Distribution of lead deviations across all situations.

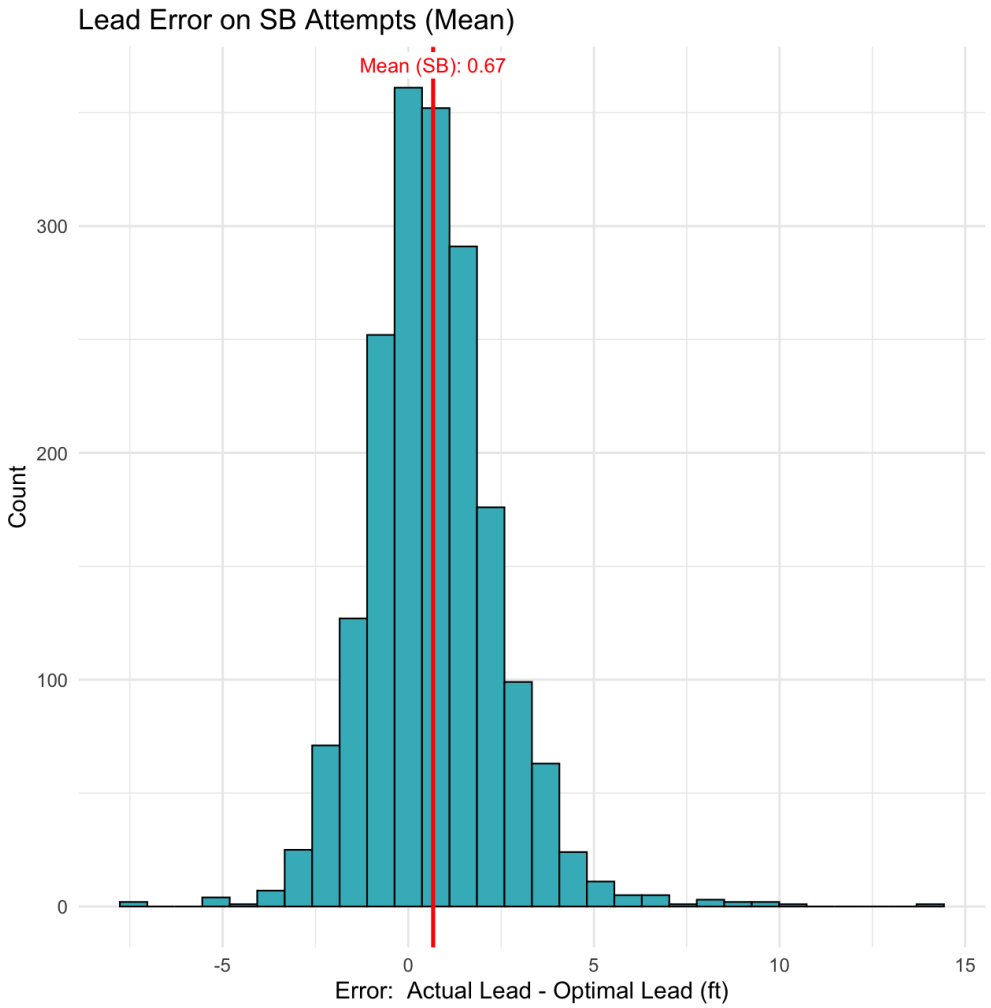


Figure 2: Distribution of lead deviations when a stolen base is attempted. Note that this distribution has a higher mean than the distribution across all situations.

Figure 3: Pete Crow-Armstrong on first base, Paul Skenes pitching, Yasmani Grandal (not seen) catching. (Marquee Sports Network, 2024)

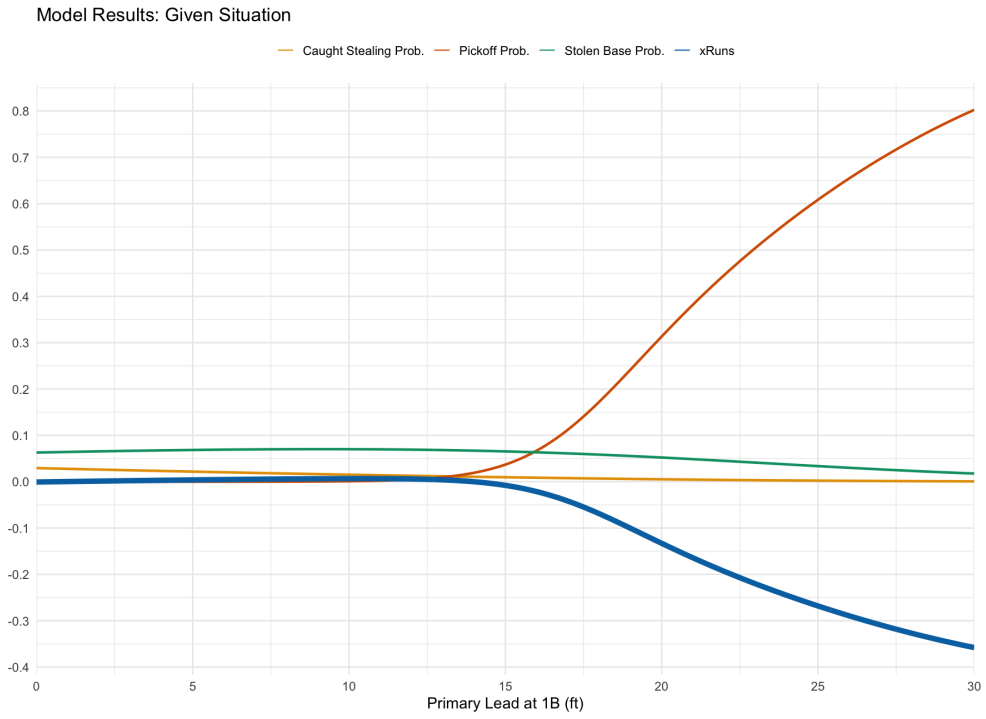


Figure 4: Our model-implied outcome probabilities and estimated xRuns as a function of Pete Crow-Armstrong's lead distance.

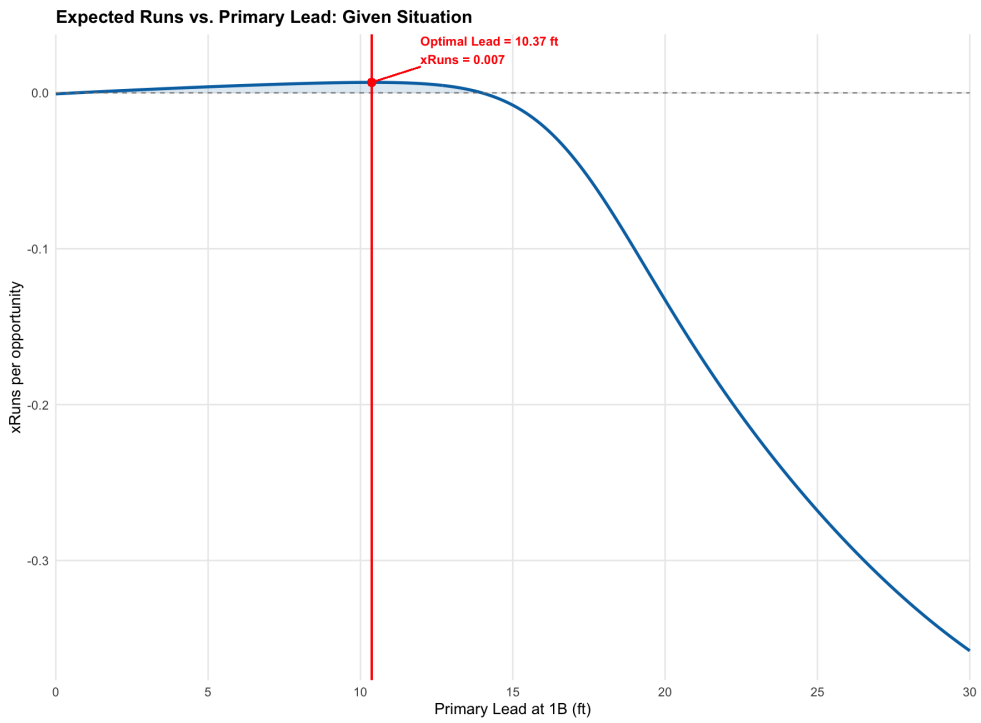


Figure 5: Expected runs as a function of Pete Crow-Armstrong's lead distance.

4 Discussion

4.1 Conclusions

This study proposes a tractable framework for selecting a primary lead off first base that *maximizes expected runs* in the base state with a runner on first and second/third empty. We model the baserunning sequence with nested logistic regressions, using runner sprint speed, catcher pop time, and pitcher hold ability ("Threat"). By mapping stage probabilities to expected runs and evaluating over the observed range of lead distances, we obtain a context-specific optimal lead L^* and a diagnostic measure of deviation $(L^{\text{obs}} - L^*)$.

Our results capture the central trade-off in lead-taking: increasing lead length raises the probability of a successful steal, but also increases pickoff risk, yielding an xRuns(L) profile with a well-defined maximum. Empirically, observed leads are modestly larger than optimal on average ($\approx +0.19$ ft), with a greater departure on steal attempts ($\approx +0.67$ ft), which we believe is consistent with unobserved intent to steal.

These findings offer an interpretable, data-driven basis for calibrating lead size that connects directly to run expectancy, recovering intuitive effects and yielding stable recommendations within the observed support.

4.2 Limitations and Future Directions

Our analysis is not without its limitations, and each suggests a direct extension. Most importantly, we do not observe *intent to steal* on a given pitch, and lead distance L may co-move with intent. In practice, clubs know when a runner has a green light to steal, and with access to intent labels could sharpen estimates by including that as a predictor. We also restrict attention to the base state with a runner on first and second/third empty; other configurations will shift both expected-run payoffs and defensive behavior, which could be addressed by a parallel analysis. Furthermore, we map outcomes to expected runs using *context-averaged linear weights* for ease, though the true value of a steal depends on game state. Substituting in context-dependent expected run calculations from a table like Table 1 would yield situation-specific recommendations. Our model also treats each pitch as independent, though we know pitchers and runners adapt within at-bats and games. Finally, our data includes only pickoffs and a sample of called pitches, under-representing swing outcomes. A similar analysis with more extensive data (including other plausibly-important covariates like batter handedness, pitch type/location, etc.) would correct for this.

4.3 Reproducibility

All analysis code and figure scripts are available HERE. Proprietary MLB data are not publicly available.

5 Acknowledgments

We would like to thank Moneyball Academy, led by Professor Abraham Wyner and the Wharton Sports Analytics and Business Initiative (WSABI), for supporting this research. We are also grateful to Major League Baseball for providing the data. We would also like to thank Jonathan Pipping, Noah Sonnenklar, and Adam Kuechler for their generous mentorship and guidance, and our teammates, Will Diflorio and Jackson Hubbard, for their contributions over the summer.

References

Baseball-Reference (n.d.). Major league baseball 1950 season summary. Sports Reference LLC.

Baseball Savant (n.d.). Base stealing run value leaderboard. Major League Baseball Advanced Media.

FanGraphs (n.d.). Re24 (run expectancy based on 24 base/out states). FanGraphs Library.

James, B. (2023). Bill james handbook excerpt: Baserunning in its own self. Sports Info Solutions.

Kagan, D. (2013). Stolen base physics. *The Physics Teacher*, 51(5):269–271.

Lindbergh, B. (2015). Statcast baserunning data and ichiro suzuki's secrets. Grantland.

Marquee Sports Network (2024). Chicago cubs vs. pittsburgh pirates [tv broadcast]. Television broadcast, Marquee Sports Network, August 28, 2024.

McMurray, J. (2015). Stolen bases in the deadball era: A relentless approach. Society for American Baseball Research.

Turocy, T. L. (2014). The economics of theft: An analysis of base stealing in baseball as a two-player game. Working paper.

Vazzana, A. (2016). Maury wills and the value of a stolen base. Society for American Baseball Research.

A Additional Lead Deviation Visualizations

Figure 6 reports the *mean absolute lead deviation* by team across all runner-on-first events. Lower values indicate closer adherence to the model's optimal lead. In our sample, the Philadelphia Phillies and St. Louis Cardinals exhibit the largest mean absolute deviations, whereas the Minnesota Twins, Detroit Tigers, and Atlanta Braves cluster near zero (closest to optimal).

Figure 7 displays the runners with the smallest *absolute* mean lead deviation (i.e., most consistently near-optimal). Figure 8 shows the tails of the signed distribution: panel (a) lists the most *aggressive* runners (positive mean deviation; leads longer than optimal) and panel (b) lists the most *conservative* runners (negative mean deviation; leads shorter than optimal). Jazz Chisholm Jr., Michael Conforto, and Dansby Swanson obey our model most closely (on average). Daniel Vogelbach, Mike Ford, and Luken Baker take the most aggressive leads on average, while Brooks Baldwin, Anthony Volpe, and Carlos Correa take the most conservative leads on average.

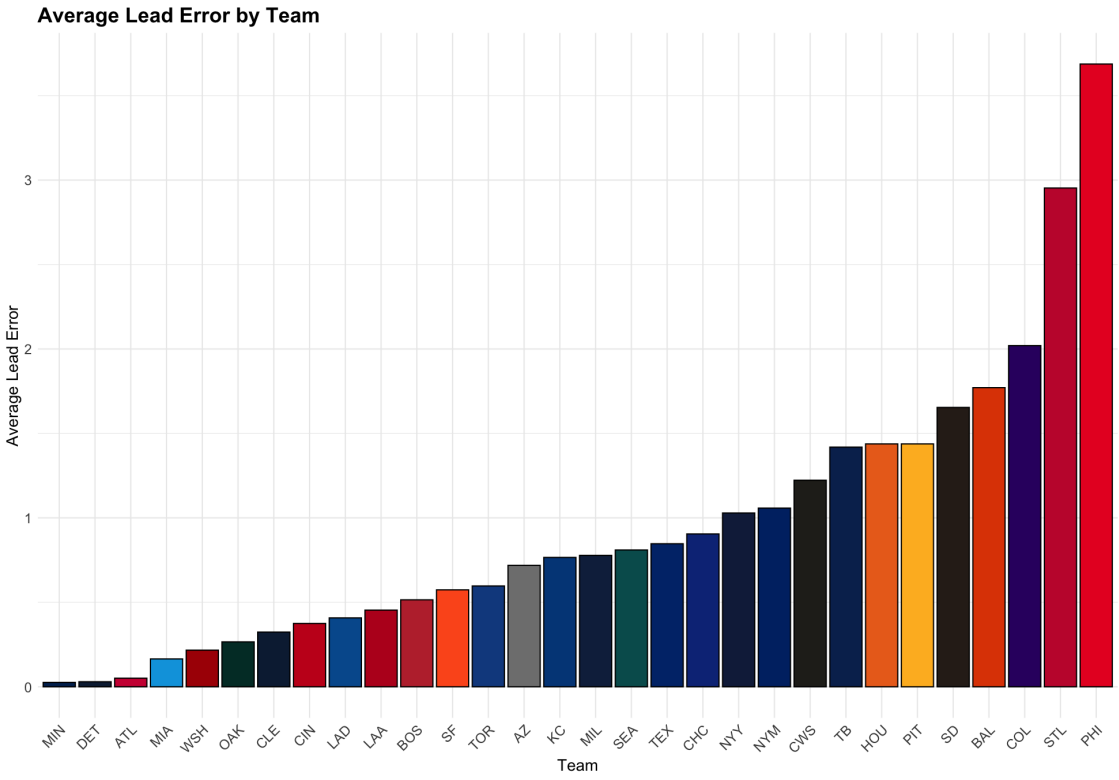


Figure 6: Bar plot of average absolute lead deviation by team.

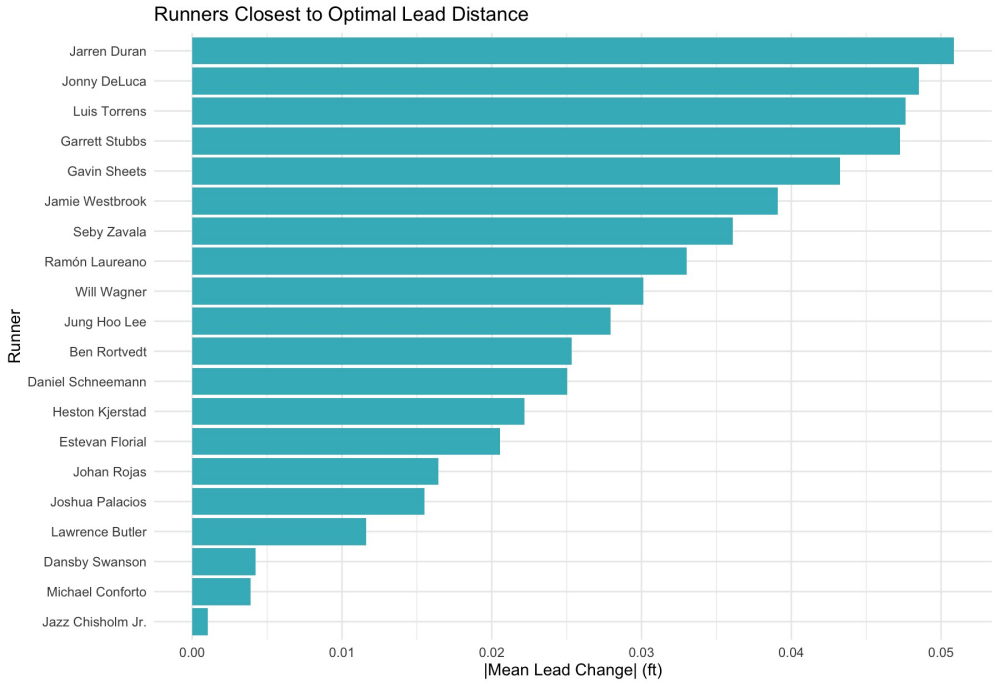
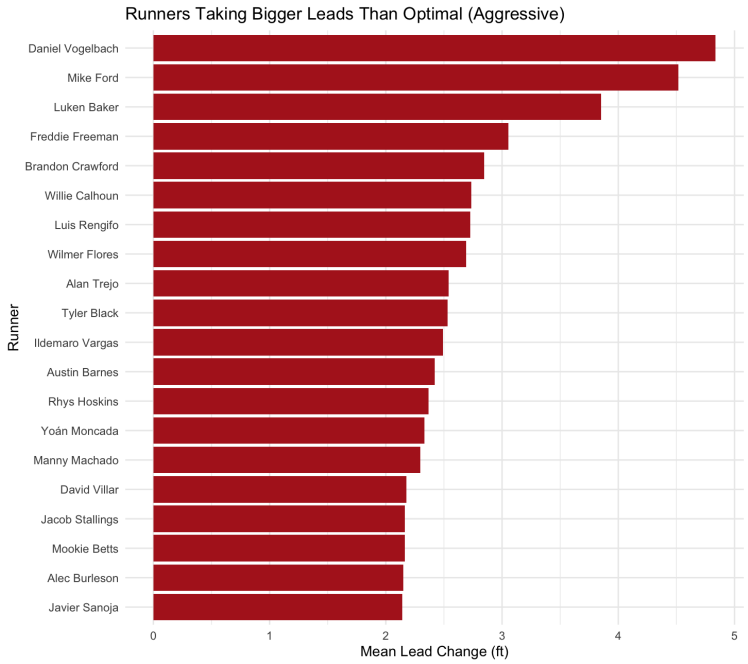
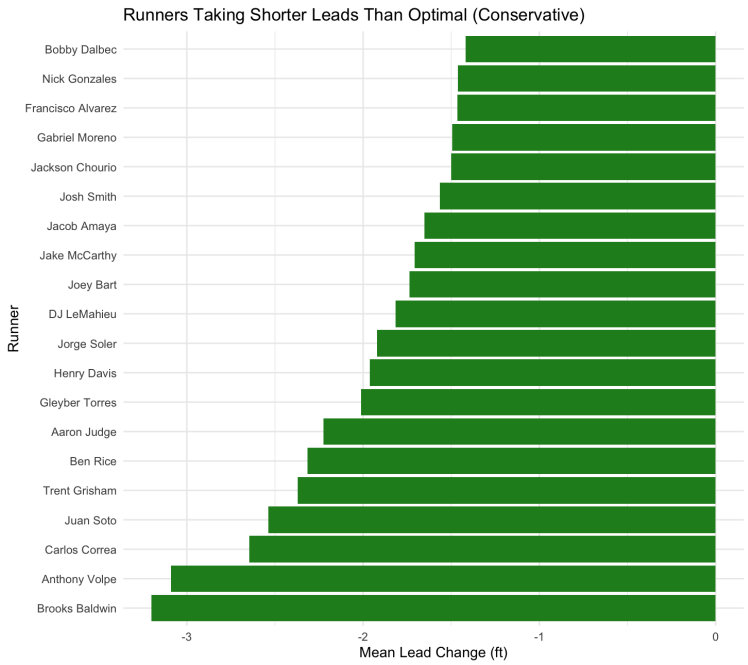


Figure 7: Baserunners with the smallest absolute average lead deviation.



(a) Baserunners with the most positive average lead deviation.



(b) Baserunners with the most negative average lead deviation.

Figure 8: Baserunners with the largest signed lead deviations: (a) aggressive (longer-than-optimal) and (b) conservative (shorter-than-optimal).