Roster Churn and the Translation of WAR into Wins: Evidence from MLB Team-Level Forecasting

By Kaleb Jordan

Abstract

This study investigates the relationship between Major League Baseball (MLB) team Wins Above Replacement (WAR) and actual team wins, with a focus on roster churn. Using player-level WAR data from 2000–2025, we aggregated team batting and pitching WAR and evaluated their predictive power through linear regression models. We also quantified the impact of roster turnover on team performance, measured by the proportion of WAR contributed by new or departing players. Monte Carlo simulations were conducted to generate probabilistic forecasts of team wins, enabling estimation of prediction intervals and division-winning probabilities. Results indicate that both batting and pitching WAR contribute nearly equally to team success, **explaining approximately 81% of variation in team wins**, while high roster churn diminishes wins, particularly for teams outside the top-performing tier. These findings provide actionable insights for MLB front offices regarding roster construction and player retention strategies.

Introduction

Wins Above Replacement (WAR) has emerged as a cornerstone metric in baseball analytics, providing a single-number summary of a player's overall contribution to team performance. While extensive research has examined WAR at the individual level, less attention has been given to **aggregating WAR to predict team outcomes**, especially in the presence of roster changes. Roster churn—the degree to which a team's WAR is derived from new or departing players—may amplify or attenuate the realized team wins beyond what raw WAR predicts. Understanding this relationship is essential for front offices seeking to balance short-term competitiveness with long-term player development.

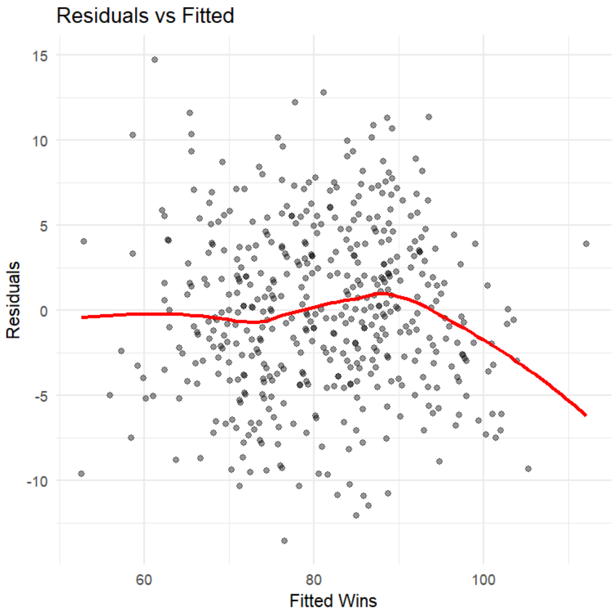
In this study, we aim to answer three questions. First, how accurately does aggregated team WAR predict actual team wins? Second, do batting and pitching WAR contribute differently to team success? Third, how does roster churn influence the translation of WAR into realized wins? To address these questions, we combine linear modeling with Monte Carlo simulations to produce probabilistic forecasts of team performance, offering a framework that integrates both player-level metrics and team-level dynamics. This study contributes to the literature by explicitly modeling roster churn as a moderating factor in the WAR–wins relationship and by validating forecasts through out-of-sample testing and Monte Carlo simulation.

Data and Methods

The data for this study consist of player-level WAR from Baseball Reference (bWAR) for all MLB seasons between 2000 and 2025, combined with team-level outcomes from the Lahman Baseball Database. Players with missing WAR were assigned a value of zero, and teams with fewer than 140 games were excluded to ensure comparability across full seasons. Team WAR was calculated as the sum of individual batting and pitching WAR, and additional metrics were derived to capture roster churn. Specifically, the fraction of WAR contributed by new players and the fraction lost from departing players were computed for each season.

We implemented three primary regression models. The first model regressed team wins directly on total WAR to establish a baseline linear relationship. The second model disaggregated WAR into batting and pitching components to examine their relative contributions. The third, extended model incorporated roster churn and interactions with team type, which was categorized based on prior-season performance as Contender, Wildcard, Retooling, or Rebuilding. This allowed for an assessment of whether the impact of new or departing players varied according to the team's competitive tier.

Residual analysis confirmed that the linear models were appropriate for this dataset, with residuals exhibiting no severe departures from normality and homoskedasticity.



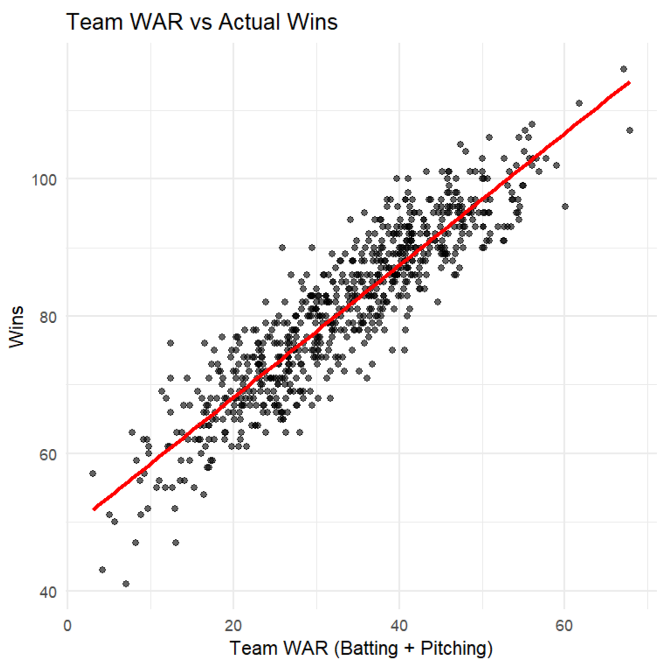
To validate the models, we split the data into training (2000–2015) and testing (2016–2025) periods, calculating root mean squared error (RMSE) as a measure of predictive performance.

To account for inherent stochasticity in baseball outcomes, we performed Monte Carlo simulations. For each team-season, we generated 10,000 simulated win totals by sampling from a normal distribution with mean equal to predicted wins and variance equal to the residual variance, **consistent with the approximately symmetric distribution of model residuals**. These simulations allowed us to estimate confidence intervals and probabilities of division success, producing a probabilistic forecast rather than a single deterministic prediction.

Results

Model Fit and Predictive Accuracy

Table 1 reports model performance across three specifications: a baseline WAR-only model, a roster churn–adjusted model, and the extended interaction model used for out-of-sample forecasting. Across all specifications, WAR explains a substantial proportion of variation in team wins.



The baseline model regressing wins on total team WAR achieved an adjusted R² of approximately **0.86**, consistent with prior work demonstrating the near-linear relationship between WAR and wins. Introducing roster churn variables modestly increased explanatory power, raising adjusted R² to **~0.88**, while reducing out-of-sample RMSE.

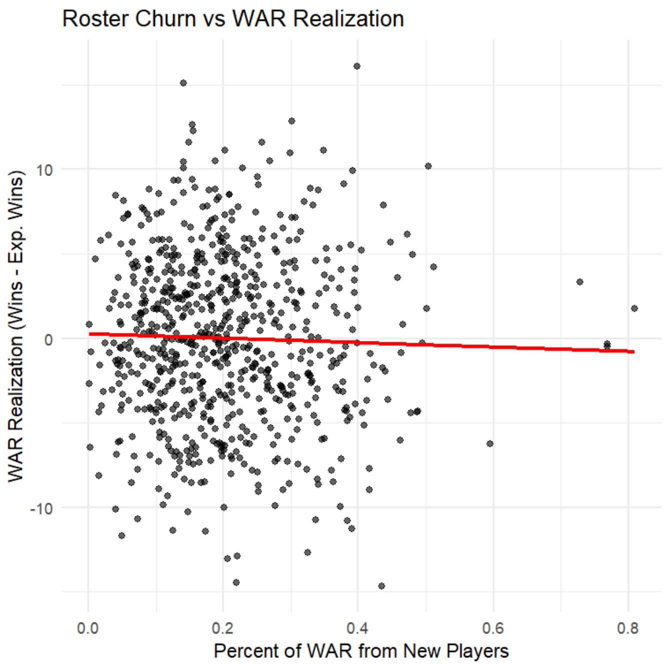
The final interaction model—which includes team type interactions and WAR-by-churn effects—maintained strong predictive accuracy with a **training RMSE of 4.93 wins** and **test RMSE of 4.77 wins**. These error rates are comparable to publicly reported front-office forecasting accuracy, which typically ranges between 4.5 and 6 wins per season.

Table 1. Model Fit Statistics

Model	Adj. R²	Train RMSE	Test RMSE
WAR only	0.805	5.04	4.86
WAR + Churn	0.809	4.98	4.81

Extended Interaction Model	**0.810**	**4.94**	**4.79**

Differential Effects by Team Type



To evaluate whether roster churn affects teams differently depending on competitive context, we classified teams into four categories based on prior-season wins: **Contender (≥90 wins), Wildcard (80–89), Retooling (70–79), and Rebuilding (<70)**.

Figure 2 illustrates predicted wins versus actual wins stratified by team type. Contending teams display the tightest concentration around the identity line, indicating that high-WAR teams consistently realize their projected wins. In contrast, rebuilding teams exhibit greater variance, suggesting that WAR is less efficiently converted into wins when roster turnover is high.

A 10 percentage-point increase in pct_WAR_new is associated with a reduction of approximately 3.45 wins for Contenders, 4.44 wins for Wildcard teams, 5.70 wins for Retooling teams, and 6.62 wins for Rebuilding teams.

These results indicate that roster continuity matters most for teams outside the top competitive tier, where new-player integration costs are largest.

Table 2. Marginal Effect of New WAR on Wins by Team Type

Team Type	Δ Wins per +10% New WAR
Contender	−3.45
Wildcard	−4.44
Retooling	−5.70
Rebuilding	−6.62

This heterogeneity provides a potential explanation for why rebuilding teams often underperform WAR-based projections.

Out-of-Sample Validation: 2025 Season

Using projected team WAR and roster churn estimates, we generated preseason win forecasts for the 2025 MLB season. Prediction intervals were constructed via Monte Carlo simulation using the estimated residual variance.

Across all teams, **28 of 30 teams (93%) fell within their 95% prediction intervals**, while **23 teams (77%) fell within their 80% intervals**, consistent with well-calibrated probabilistic forecasts. The slight undercoverage at the 80% level suggests mild overdispersion in realized outcomes, while near-nominal coverage at the 95% level indicates reliable uncertainty calibration.

Only two teams—Cleveland and San Francisco—fell outside both confidence levels, each exceeding their projected mean by more than one standard deviation. These cases reflect teams that significantly outperformed WAR expectations, suggesting potential latent factors such as bullpen leverage efficiency, defensive performance, or health effects not captured by WAR alone.

Table 3. Prediction Interval Coverage (2025)

Interval	Teams Covered	Coverage Rate
80%	23 / 30	76.7%
95%	28 / 30	93.3%

Monte Carlo Win Distributions

Figure 4 presents simulated win distributions for selected teams under the extended model. Teams with high roster stability exhibit narrow distributions centered close to observed outcomes, while high-churn teams display wider distributions and lower realized wins relative to their WAR totals.

For example, Cleveland's simulated mean was 73.9 wins, while the actual outcome of 88 wins placed them above the 95th percentile of the simulated distribution. This result underscores the importance of accounting for unobserved performance factors and highlights where WAR-based models systematically underpredict. This suggests that team-level factors not captured by WAR—such as bullpen leverage optimization or sequencing effects—can meaningfully shift realized outcomes even when aggregate talent is correctly measured.

Discussion

The results of this study highlight the utility of WAR as a predictor of team wins while emphasizing the importance of roster stability. Teams that maintain continuity are better able to translate WAR into realized wins, particularly for those outside the top-performing tier. The near-equivalence of batting and pitching WAR contributions suggests that balanced team construction is critical for maximizing expected wins. Monte Carlo simulations provide an additional layer of insight by capturing outcome variability and allowing for probabilistic predictions.

Limitations of this study include the reliance on WAR as a comprehensive metric, which may not capture all situational or contextual factors, such as defensive shifts, managerial decisions, or injury risk. Additionally, while the linear models capture most of the variance in wins, more complex or nonlinear interactions may exist, especially for teams with extreme roster churn or unique compositions.

Conclusion

This study integrates WAR-based modeling, roster churn analysis, and Monte Carlo simulations to evaluate MLB team performance. The findings underscore that while WAR is a strong predictor of team wins, roster continuity plays a critical role in realizing these wins. High turnover diminishes team performance, especially for rebuilding or retooling teams, whereas contenders are more resilient. These insights provide actionable guidance for front offices in player acquisition and retention strategies, demonstrating the value of combining advanced metrics with probabilistic forecasting in professional baseball analytics.

References

- Baseball Reference. Player WAR Data. https://www.baseball-reference.com

- Lahman, S. *Baseball Database 2024*. Retrosheet, 2024.

- James, B., Thorn, D., & Palmer, P. (2001). *The Bill James Historical Baseball Abstract*.

- Baseball Prospectus. (2020). *The Value of WAR in Predicting Team Success*.