Final Project

Pedro Girardi, Matt Kaye, and Ben Schwartz

11/19/2018

Abstract

For the last two decades, Sabermetricians have been working to accurately represent player performance to provide front offices with data to be used to value a player's contribution to their team. As a result, Wins Above Replacement (WAR) has emerged as leading metric for player performance. In this study, we take a Bayesian approach to evaluating the calibration of WAR with MLB data post-1950. Hamiltonian Monte Carlo is used to fit two predictive models, one using WAR and the other using vanilla count statistics.

Keywords: Sabermetrics, Bayesian Methods, Markov Chain Monte Carlo, Poisson Regression

INTRODUCTION

Wins Above Replacement (WAR) is a statistic that has recently emerged as the best all-in-one assessment of player performance in Major League Baseball (MLB). WAR attempts to quantify the number of wins a player provides their team compared to a readily available substitute player (generally thought of as a AAA player who could be called up). This idea that a single statistic can show the value a player contributes to his team raises the question on whether or not WAR more accurately predicts wins than a combination of other individual statistics such as hits, runs batted in (RBI), total bases, stolen bases, strikeouts, and walks can. As an extension, we will look at models using WAR versus individual count statistics to predict team wins, examining whether or not one model is better for predicting the amount of wins in a given year given values for each statistic at the team level.

WAR only recently has been considered in the baseball community, and the MLB does not recognize it as an official statistic. However, there have still been studies trying to relate WAR to team wins because the baseball community broadly uses WAR as its main evaluative statistic. In a 2009 blog post from David Cameron, it was found that in the 2009 season, the correlation between a team's wins and WAR was .83. More recently, Glenn DuPaul had a blog post on *The Hardball Times* examining WAR's ability to be used as a predictive statistic for wins. He found that the correlation coefficients for in-season wins related to aggregate team WAR was .91. However, when DuPaul attempted to predict wins in a future season using WAR, he discovered that the correlation between WAR and wins was only .59. In his test, he only used thirty teams' statistics, randomly sampled from the 2007-2011 seasons. In our project, we will attempt to fit a model using data from the 1950-2018 seasons.

To look at the question of using WAR versus individual statistics to predict team wins, we will want to examine the question how well WAR is calibrated for in-season performance. To do this, we will be fitting models with data from the 1950 through current seasons with both WAR and count statistics.

LITERATURE REVIEW AND INSTITUTIONAL DETAILS

In 1974, Gerald Scully published a paper on player compensation in Major League Baseball that has become foundational in sports economics. He argued that player salaries are a function of player performance, and that player performance is a function slugging percentage (SLG) for hitters and strikeout-to-walk ratio (K/BB) for pitchers. His use of SLG and K/BB were one of the first uses of what has become known as Sabermetrics: the practice of using advanced statistics to measure player performance. Scully's paper was followed up by Anthony Krautmann in 1999, who substituted earned run average (ERA) for K/BB, and added a second model that used total bases (TB) in place of SLG. Fast forward to today, and Sabermetrics has exploded. In 2018's MLB, every team has an analytics department whose primary goal is to determine how much money a player should be paid. In doing this, they must do as Scully and Krautmann did and attempt to evaluate player performance in order to put a dollar value on a player that they may sign.

Nowadays, front offices do not use SLG, K/BB, ERA, or TB to evaluate players. Rather, they use WAR or other advanced metrics, which have become the gold-standard evaluative statistics in the MLB. The benefit of WAR in particular, per Fangraphs, is that it is a context-, league-, and park-neutral statistic. This means that we can compare WAR across players, teams, and eras, which allows us to not only evaluate player performance for someone on one team in one year, but to compare that player to a player for a team that no longer exists from fifty years ago, for example.

This raises the question of whether or not WAR has been successful, and, especially, how well does WAR actually predict performance? In other words, how well does a team's total WAR in a certain year predict how well they did that year? If WAR is as well-calibrated as the Sabermetrics community seems to think it is, then it should be almost perfectly correlated with wins, and will certainly outperform other statistics like those that Scully and Krautmann used in their models of player and team performance years ago.

DATA

There are a few baseball databases that have free and easily downloadable statistics. All of the statistics that we used as the explanatory variables in the model came from Fangraphs. Fangraphs has data at the player level, so we used Dplyr to aggregate all of the player-level data up to team-level data by summing the counts of all of our variables grouped by team and year. This left us with 1,630 team-year pairs (i.e. observations in our table) for every year after 1950 that we would use to build the model.

Win totals for every season from 1950 came from Baseball Reference, which has exportable win data for every team since 1901. We opted to use data from 1950 to 2018 because the structure of the League was different in its early years from how it is today.

After downloading and aggregating the data, we joined the two data sets into one by matching the team-year pairs from both of the tables, which left us with one data frame of counting statistics and wins at the team level for each team in every year since 1950. The statistics that we have kept track of are earned runs, home runs allowed, walks allowed, strikeouts (by pitchers and hitters), total bases, runs scored, RBI, walks by hitters, WAR for hitters, WAR for pitchers, overall WAR (hitters plus pitchers) and team wins. We chose all of these statistics because they make up the majority of the basic counting statistics for major league players, are the statistics used to calculate averaged statistics like batting average, slugging percentage, and earned run average, and were statistics used in the literature referenced above. Thus, we would expect that they would account for most of the variability in team wins in a given season.

It is worth noting that we standardized all of our variables but `wins`. Our motivation to standardize the variables was several convergence issues with fitting the model with the original dataset and almost no effective draws from our Markon chains in our HMC process. Once we standardized our explanatory variables, the chains were able to converge and produce significantly more effective draws.

Table 5 in the Appendix contains summary statistics for all of our variables of interest.

METHODS

Since we are modeling a posterior with a count value, we use poisson regression with parameter λ representing the mean number of wins. In our comparison, we are looking at two models - a model considering only WAR to account for wins in a season (Model 1), as well as a model using other count statistics (Model 2).

For Model 1, we have a poisson posterior on the log number of wins being a linear function of our intercept and beta coefficient for WAR.

Model 1: WAR Model

$$log(Wins) \sim Pois(\lambda)$$

$$\log(\lambda) = \alpha + \beta \cdot WAR$$

$$\alpha \sim N(4, 3)$$

$$\beta \sim N(0, .5)$$

For our prior specifications, we took into account the previous studies about WAR and how it is calibrated for in season performance. Since the previous studies found wins were represented as a linear function with WAR as $Wins = 52 + 1.0 \cdot WAR$ (i.e., an additional unit of WAR increases the expected number of wins by 1), we wanted to incorporate this knowledge into our prior specification. As a result, we put a normal prior on our intercept coefficient, centered at 4 with a standard deviation of 3, which is rather diffuse considering we expect our intercept to be close to $e^4 = 54.6$. Our slope coefficient for β is best represented around 0, as we would expect a small multiplicative change in wins for every additional increase of 10.8 (sample standard deviation of WAR) in WAR. Recall our variables are scaled, so e^β actually represents the expected multiplicative change in wins given an increase of $sd(WAR)$ in WAR.

Model 2 accounts for several aggregated count statistics for a team. Since WAR is calculated as a function of count statistics and other adjustments, it should be able to predict wins more accurately than a model taking into account only count statistics. Model 2 uses a team's earned runs, pitching walks, strikeouts (pitcher), total bases, runs, stolen bases, walks (hitters), strikeouts (hitters) to predict wins.

For the same reason as in Model 1, we use Poisson regression to model wins because our response variable (wins) is a count. More thought was needed in our prior specification for this model. We wanted our priors to be regularizing in order to reduce overfitting our model. However, we could not make them as informative as the prior on WAR in Model 1 because the theoretical relationship between hits, for example, and wins is not as clear. However, we still believe that the multiplicative change associated with any of these counting statistics should be very close to zero. The rationale behind this choice of a prior is straightforward: a unit-increase in the standard deviation of something like runs (i.e. a team scoring about 108 more runs over the course of a season) should have a very small multiplicative effect on that team's total wins, because, in real terms, a multilicative change of 1.1 could correspond to as many as 10 wins. We would not expect, for instance, a team with additional 108 runs to increase the number of wins by 40%. As such, we again centered our priors at 0 with a very small standard deviation of .1.

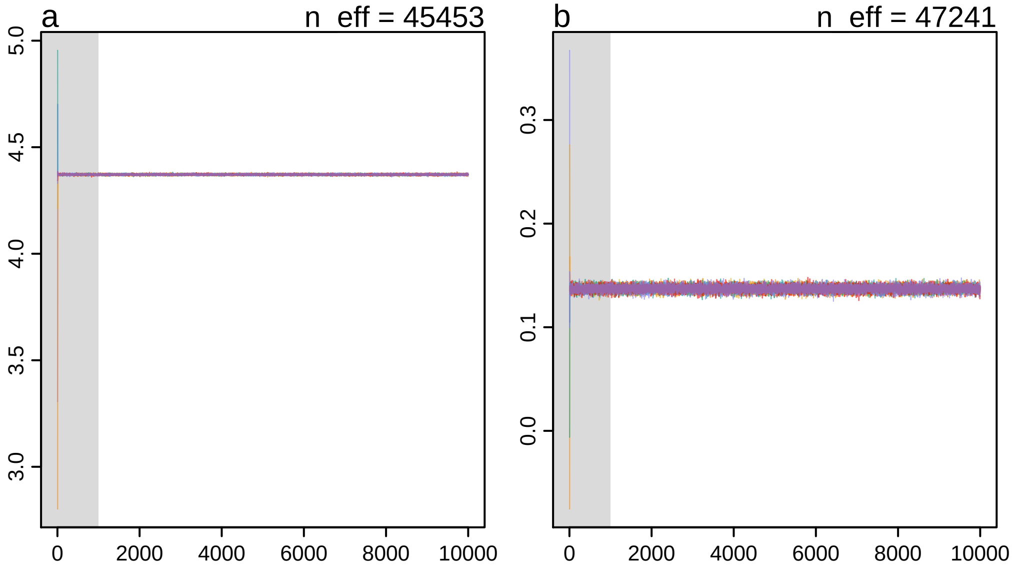
Model 2: Count Statistics

$$\log(Wins) \sim Pois(\lambda)$$
$$\lambda = \alpha + \beta_1(ER) + \beta_2(BB_p) + \beta_3(K_p) +$$
$$\beta_4(HR) + \beta_5(SB) + \beta_6(BB_h) + \beta_7(K_h) + \beta_8(TB)$$
$$\alpha \sim N(4,3)$$
$$\beta_{1,\ldots,8} \sim N(0,.1)$$

RESULTS

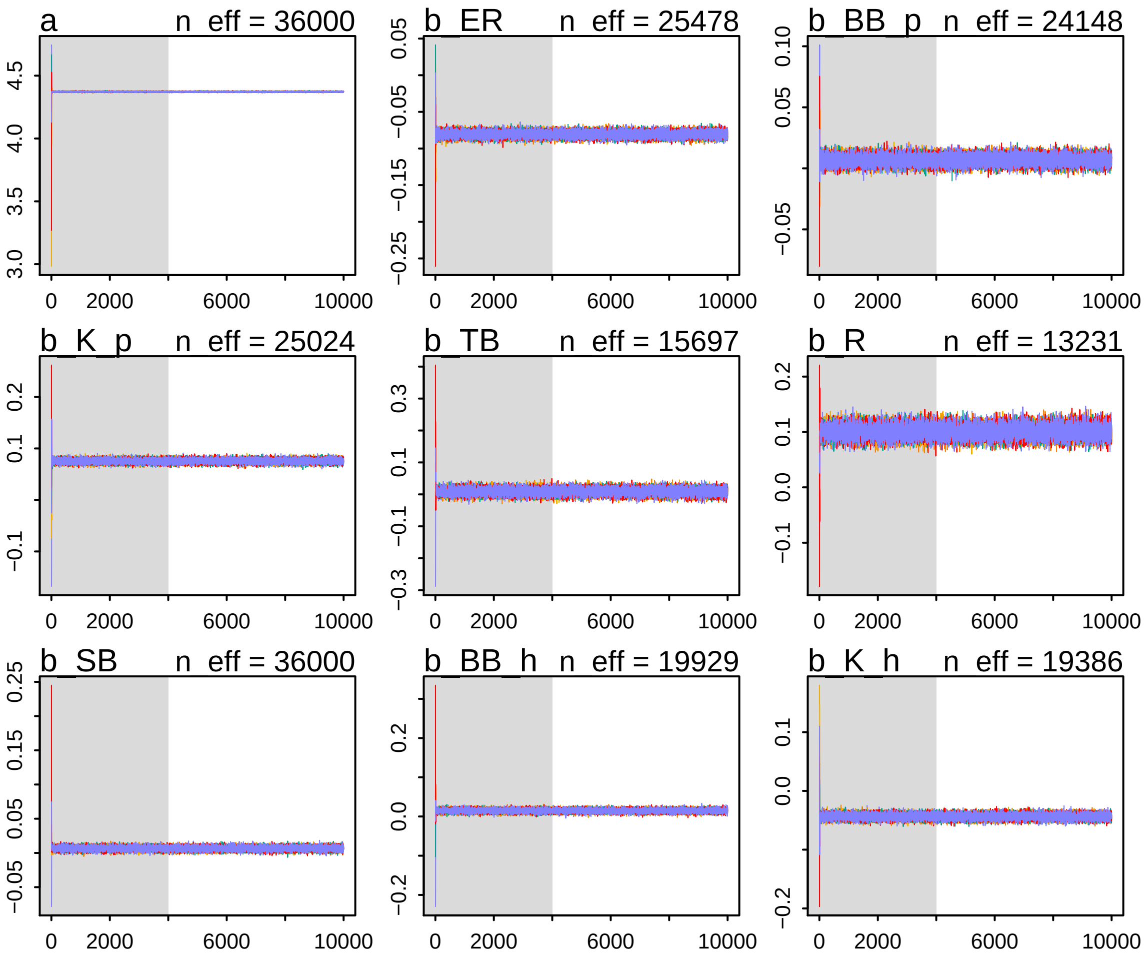
We used Hamiltonian Monte Carlo (HMC) to estimate both of our Poisson models. For Model 1, we used a warmup period of 1000 iterations, 6 chains, and 10,000 iterations overall. For Model 2, we still used 10,000 iterations, but we used a warmup period of 4,000 instead of 1,000 with 6 chains. Using 6 chains and 10,000 iterations let us draw large numbers of effective samples for each predictor in our model.

Figure 1: Model 1 HMC Simulation



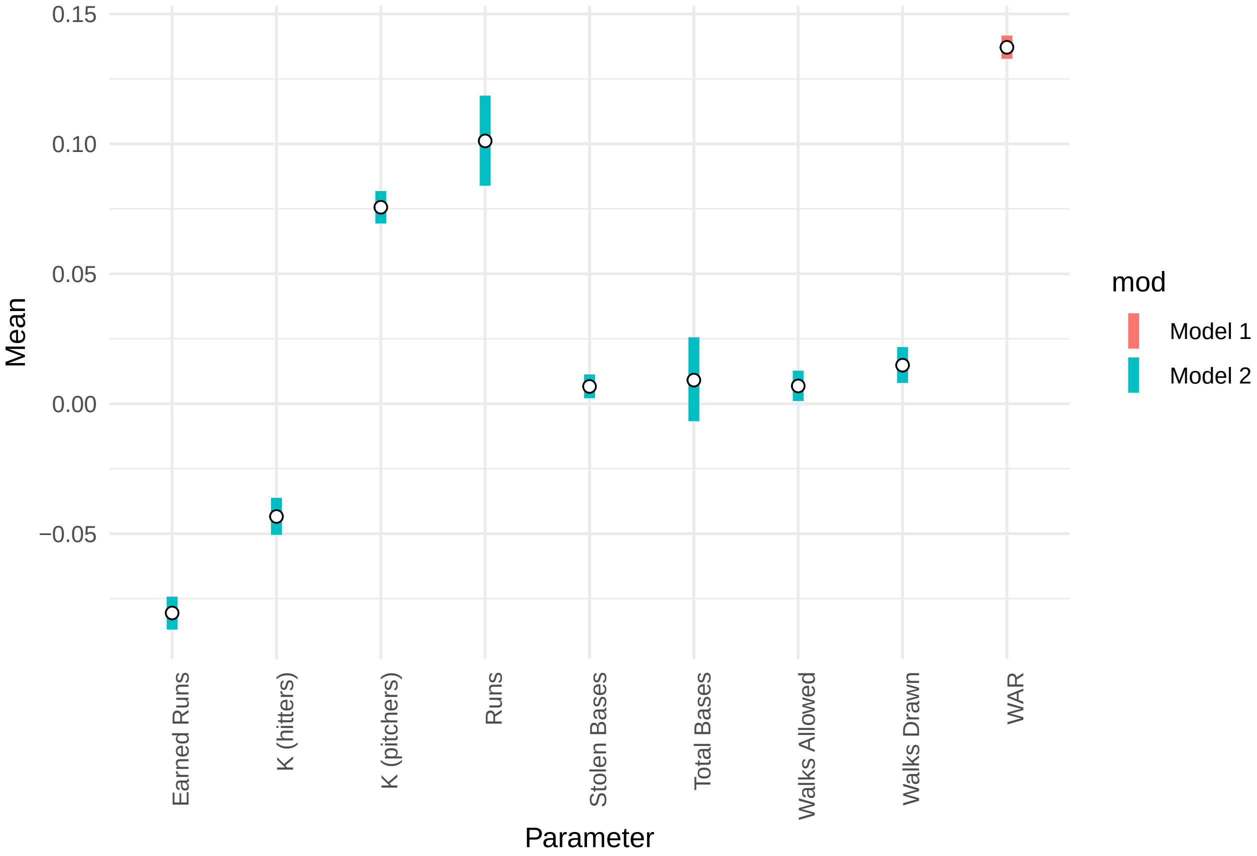
Model 1 quickly converged and stabilized. Analyzing Figure 1, one sees that, for both *a* and *b*, the chains started at different locations but converged to the same distribution at around 4.4 and 0.15, respectively, after relatively few iterations. Furthermore, we observe that both parameters had over 45,000 effective draws. Therefore, both the convergence to a single distribution and the high number of effective draws indicate well-behaved chains.

Figure 2: Model 2 HMC Simulation



Although Model 2 had eight predictors and therefore increased complexity, we still observed very well behaved chains. For all the parameters, the chains quickly converged to their stationairy distributions. The number of effective draws was above 13,000 for every parameter, which is a good indication that we are getting an accurate representation of our posterior.

Figure 3: Model Parameter Estimates and Confidence Intervals

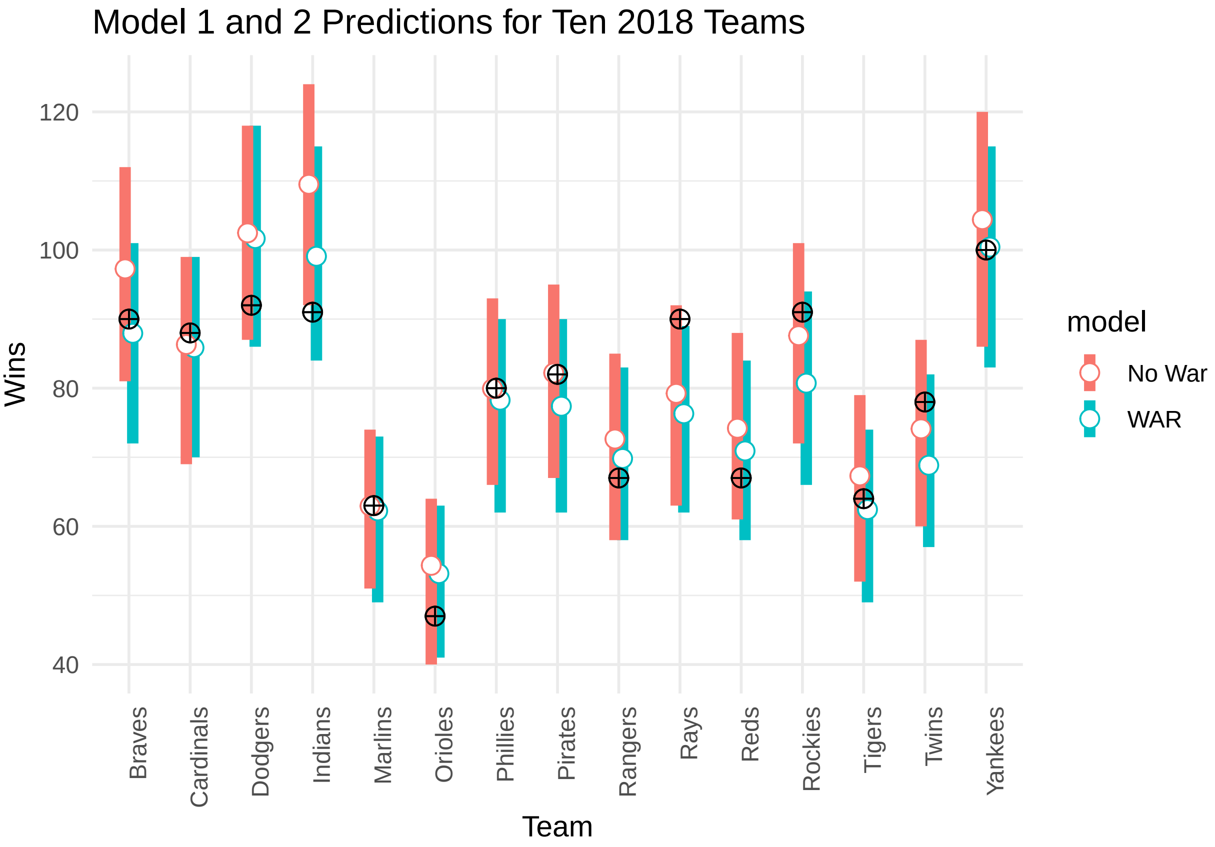


If we look at the above plot, one general thing to notice is that any parameter estimate which is contained above zero will result in an predicted increase in wins for an increase in the predictor variable's value. Any parameter which is contained below zero will result in a predicted decrease in wins for an increase in the predictor variable's value. For instance, it is interesting to note that in these models, the largest factors for a team's success other than WAR seem to be runs given up and runs scored. This makes sense, as in the most simplistic way, if your team can score more runs than your opponent, you should also be able to win more games than you lose.

Table 1: WAIC Comparison of Models

Model	WAIC	pWAIC	dWAIC	weight	SE	dSE
Model 1	10972.58	1.095	0.000	1	36.837	NA
Model 2	11115.01	5.438	142.432	0	39.909	32.94

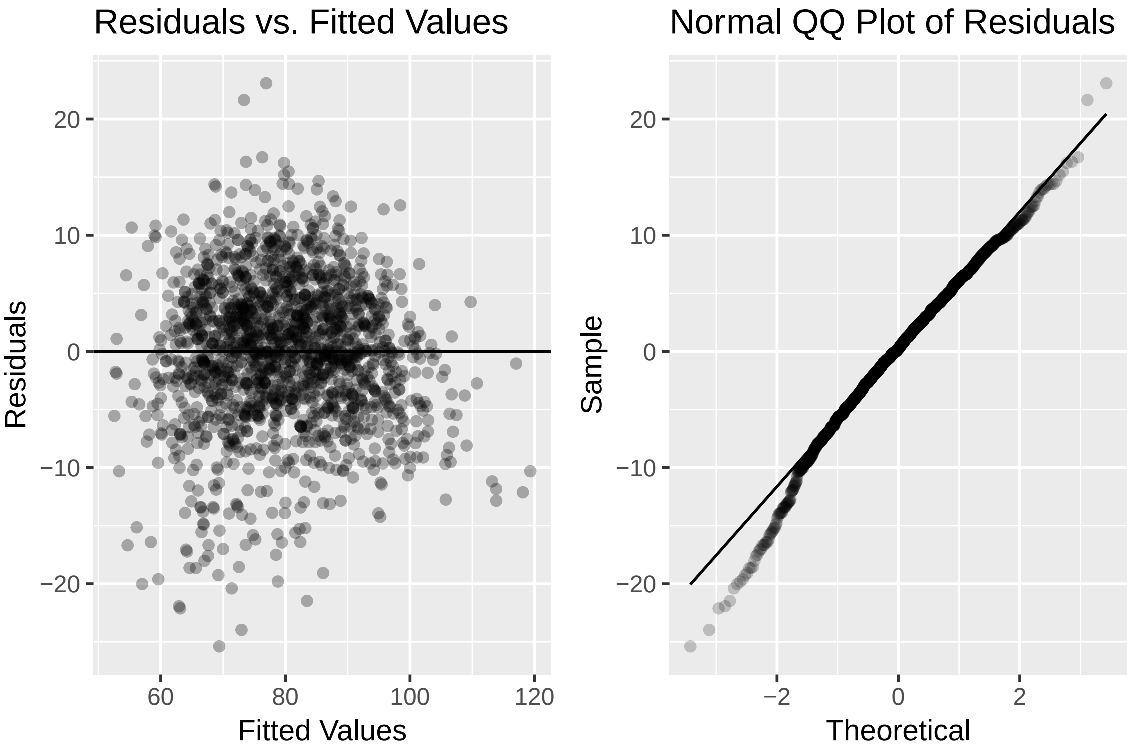
The table above shows a comparison of the two models we fit. Model 1, unsurprisingly, performs much better than Model 2. The difference in WAIC between Model 2 and Model 1 is 142.432, which is a large difference (large enough to make Model 1's Akaike Weight 1). Another way of comparing the accuracy of both models is through prediction plots. The plot below shows the different predictions and 89% HPDI prediction intervals for both models. As we can see, Model 1 has both tighter prediction intervals and generally more accurate point predictions than Model 2.



In Model 1, the exponent of the estimated intercept indicates the expected number of wins for a team that has the average WAR. The average WAR in our dataset is 31.898, and the expected number of wins for a team with average WAR is $e^{\hat{\alpha}} = 79.18$. The coefficient of β is 0.137, and it indicates that, for every unit increase in WAR, there is an increase in the expected number of wins by a factor of 1.013.

The coefficients in Model 2 has similar interpretations. The intercept indicates that the expected number of wins for a team with league average values for all predictor variables has an expected number of wins of 79.218. Because there is a large number of parameters in the model, we will interpret only two of them. β_{ER}, the coefficient for Earned Runs, and β_{SB}, the coefficient for Stolen Bases. Our model estimated $\hat{\beta}_{ER} = -0.081$, so the expected multiplicative change in the number of wins given a one standard deviation in Earned Runs (which corresponds to 89.693 earned runs) is 0.923. This makes sense, since we would expect the number of wins to decrease as a team allows more runs. For stolen bases, our model estimated $\hat{\beta}_{SB} = 0.007$, so the expected multiplicative change in the number of wins given a one standard deviation change in stolen bases is 1.007. One standard deviation for stolen bases is 40.686. This result also makes sense, since one would expect the number of wins to increase as a team steals more bases.

Finally, we needed to check for violations of basic regression assumptions. The plots below show residuals against fitted values and a quantile-quantile plot of our residuals to check for normality.



Both of these diagnostic plots look good. The residuals seem to be randomly scattered about zero with no pattern to them, so Model 1 is not violating the heteroskedasticity assumption, the linearity assumption, or the autocorrelation assumption. Additionally, since there is only one predictor, we know that there is no collinearity in the predictors in Model 1. Finally, although there seems to be a slight left skew in the residuals, our sample size of 1,630 observations is big enough that we should not worry about potentially violating the normality assumption for the residuals.

Conclusions

Throughout this study, it has been interesting to see the effects of certain individual count statistics when comparing them to the predicted amount of wins in our second model. However, our first model is by far the more accurate model in terms of accurately predicting wins. While WAR is not perfect for predicting wins, the general accuracy of the model indicates that WAR is a very well calibrated statistic, as expected.

Going forward, it would be equally interesting to test how well basic averaging stats, including batting average, slugging percentage, and more advance averaging and count statistics, including wRC+ and BABIP can predict wins, and whether or not they do a better job than the counting statistics that we chose. Additionally, we would like to be able to create a model which can test year-to-year prediction ability using WAR by fitting a predictive model using WAR in a previous season to predict wins in a future season after accounting for age, trades and signings, and other random noise.

There are clear implications to models like these for front offices. If a team's ultimate goal is to maximize its win totals in an effort to win the World Series, then they should definitely care about how well calibrated their statistics are that they are using to measure and predict performance, in order to give themselves the best possible chance.

Appendix

Table 2: Model 1

Parameter	Mean	StdDev	lower 0.89	upper 0.89	n_eff	Rhat
Intercept	4.372	0.003	4.367	4.376	45452.96	1
WAR	0.137	0.003	0.133	0.142	47241.02	1

Table 3: Model 2

Parameter	Mean	StdDev	lower 0.89	upper 0.89	n_eff	Rhat
Intercept	4.372	0.003	4.368	4.377	36000.00	1
Earned Runs	-0.081	0.004	-0.087	-0.074	25478.00	1
Walks Allowed	0.007	0.004	0.001	0.013	24148.40	1
K (pitchers)	0.076	0.004	0.069	0.082	25024.13	1
Total Bases	0.009	0.010	-0.007	0.026	15696.86	1
Runs	0.101	0.011	0.084	0.119	13231.24	1
Stolen Bases	0.007	0.003	0.002	0.011	36000.00	1
Walks Drawn	0.015	0.004	0.008	0.022	19928.76	1
K (hitters)	-0.043	0.004	-0.050	-0.036	19385.73	1

Table 4: Parameter Longnames

parameter	longname
a	Intercept
b	WAR
b_ER	Earned Runs
b_BB_p	Walks Allowed
b_K_p	Strikeouts (Pitchers)
b_TB	Total Bases
b_R	Runs
b_SB	Stolen Bases
b_BB_h	Walks Drawn
b_K_h	Strikeouts (Hitters)

Table 5: Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
Wins	1,630	79.924	12.476	37	72	81	89	116
Earned Runs	1,630	574.751	89.693	251	515	573	632.8	920
Walks Allowed	1,630	475.372	74.229	248	426.2	473	523	770
K (Pitchers)	1,630	881.427	198.829	341	745	871.5	1,006.8	1,605
Runs	1,630	656.593	107.515	315	587	661	730	991
Walks Drawn	1,630	485.603	82.712	244	431.2	484	538	727
K (Hitters)	1,630	883.630	193.269	346	761	875	1,000.5	1,523
Stolen Bases	1,630	88.134	40.686	15	58	83	111	340
WAR	1,630	31.898	10.884	−0.400	24.125	32.200	40.000	65.200
Total Bases	1,630	2,015.134	273.827	1,061	1,846	2,037.5	2,206	2,705

Bibliography

Carleton, R. (2012, October 2). *Baseball Therapy: WARP for People Who Didn't Like Math Class.* Retrieved November 18, 2018, from https://www.baseballprospectus.com/news/article/18511/baseball-therapy-warp-for-people-who-didnt-like-math-class/

Carleton, R. (2013, September 11). *Reworking WARP: Why We Need Replacement Level.* Retrieved November 18, 2018, from https://www.baseballprospectus.com/news/article/21773/reworking-warp-why-we-need-replacement-level/

DuPaul, G. (n.d.). *What is WAR good for?* Retrieved November 18, 2018, from https://www.fangraphs.com/tht/what-is-war-good-for/

Arthur, R. (2017, April 18). Do MLB Teams Undervalue Defense — Or Just Value It Differently? Retrieved November 19, 2018, from https://fivethirtyeight.com/features/do-mlb-teams-undervalue-defense-or-just-value-it-differently/

Krautmann, A. C. (1999). What's Wrong with Scully-Estimates of a Player's Marginal Revenue Product. Economic Inquiry, 37(2), 369–381. https://doi.org/10.1111/j.1465-7295.1999.tb01435.x

Scully, G. W. (1974). Pay and Performance in Major League Baseball. American Economic Review, 64(6), 915–930.

Tango, T. (2009, January 26). Misconceptions of WAR. Retrieved November 19, 2018, from http://www.insidethebook.com/ee/index.php/site/article/misconceptions_of_war/

What is WAR? FanGraphs Sabermetrics Library. (n.d.). Retrieved November 19, 2018, from https://www.fangraphs.com/library/misc/war/

Baseball Reference. (2018). *Major League Baseball Team Win Totals.* Retrieved November 15, 2018 from https://www.baseball-reference.com/leagues/MLB/index.shtml

FanGraphs. (2018). *Major League Leaderboards 1950-2018 Batters.* Retrieved November 15, 2018 from https://www.fangraphs.com/leaders.aspx?pos=all&stats=bat&lg=all&qual=y&type=8&season=2018&month=0&season1=1950&ind=1&team=0,to&rost=&age=&filter=&players=

FanGraphs. (2018). *Major League Leaderboards 1950-2018 Pitchers.* Retrieved November 15, 2018 from https://www.fangraphs.com/leaders.aspx?pos=all&stats=pit&lg=all&qual=0&type=8&season=2018&month=0&season1=1950&ind=1&team=0,to&rost=0&age=0&filter=&players=0