The paradoxical predictive relationship of wins by fWAR

Daniel Eck

11/11/2019

Summary of results

In this analysis we show that in any given season, a team's total fangraphs wins above replacement (fWAR) is a great single predictor for a team's total wins. We also show that a significant number of franchises consistently underperform or overperform their fWAR. These two facts are in direct contrast of each other and it is surprising that they can hold simultaneously. The variability of wins that is not explained by variability in fWAR may be explained by team level factors which hold for large portions of the team's existence. Explanations of this phenomenon are not provided in this analysis and are an interesting open question.

Details of the analysis follow in the next sections. Note that this analysis is not fully reproducible, the dataset that I am using was scraped from Fangraphs and fWAR is their property. I do not think it would be wise to make my data available. On that thought, thank you Fangraphs!

Dataset

The dataset is taken is from fangraphs (thank you again Fangraphs!) and it includes total team wins and fWAR for every MLB team dating back to 1920. The year 1920 is chosen because it is considered the first year of the modern era and, although fWAR can account for the run environment of a season in isolation, we do not include deadball era seasons in this analysis. The first 6 rows of the dataset are displayed below.

```
WinsvfWAR <- read.csv("WinsvfWAR.csv")
library(parallel)
library(tidyverse)
head(WinsvfWAR)
```

```
##         Team  W pWAR  hWAR fWAR year
## 1  Athletics 48 14.3 -10.3  4.0 1920
## 2     Braves 62  6.7  13.4 20.1 1920
## 3     Browns 76 10.9  27.4 38.3 1920
## 4  Cardinals 75  9.5  26.6 36.1 1920
## 5       Cubs 75 14.1  15.2 29.3 1920
## 6     Giants 86 12.2  27.0 39.2 1920
```

Analysis 1: fWAR is a good predictor of wins

To conduct this analysis we first fit the linear regression model with total team wins as the response variable and total team fWAR as the single predictor variable for every season from 1920 to 2019.

```
## split dataset by year and fit all regression models
WinsvfWARbyYear <- split(WinsvfWAR, f = as.factor(WinsvfWAR$year))
models <- mclapply(WinsvfWARbyYear, mc.cores = 12,
  function(xx){
    m <- lm(W ~ fWAR, data = xx)
```

```
    m
})
```

We also compute the coefficient of determination for each of these regression models as well as a few regression model diagnostics to assess whether or not the linear regression model is a proper tool for assessing the relationship between total team wins and total fWAR for each year. These diagnostics include a test for constant variance and normality of residuals (details are in the R code documentation).

```
## compute model diagnostics for each regression model
diagnostics <- as.data.frame(do.call(rbind, mclapply(models,
  mc.cores = 12, function(m){
    resid <- m$residuals
    fit <- m$fitted

    ## We compute the coefficient of determination for each
    ## regression model fit
    adjR2 <- summary(m)$adj.r.squared

    ## We assess constant variance by fitting the regression model
    ## with the absolute value of residuals as the response variable
    ## and regression model fitted values as the predictor. A p-value
    ## greater than 0.05 indicates constant variance (variability is
    ## not significantly changing at the 0.05 level across the
    ## predictor space).
    constvar <- summary(lm(abs(resid) ~ fit))$coefficients[2, 4] > 0.05

    ## We assess normality of residuals via a Shapiro-Wilks test for
    ## normality. A p-value greater than 0.05 indicates that the
    ## residuals follow a normal distribution (at the 0.05 level).
    normresid <- shapiro.test(resid)$p.value >= 0.05
    c(adjR2, constvar, normresid)
})))
colnames(diagnostics) <- c("adjR2", "constvar", "normality")
```

We see that a large amount of variation in total team wins is explained by the variation in total team fWAR across the seasons.

```
## coefficient of determination
summary(diagnostics$adjR2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3912  0.7658  0.8303  0.8039  0.8612  0.9451
```

Moreover, the constant variance assumption and normality of residuals assumption of these regression models hold. We see that 97% of models meet the constant variance assumption and 95% of models meet the normality of residuals assumption.

```
## check for constant variance and normality of residuals
diagnostics %>% summarise(mean(constvar), mean(normality))
```

```
##   mean(constvar) mean(normality)
## 1           0.97            0.95
```

Therefore, our regression models show that total team fWAR is a great predictor for total team wins. This finding is not that interesting in isolation, we fully expected that total team fWAR would be a great predictor of team wins in any given season. What is unexpected is that, in addition to this finding, a significant large number of teams consistently underperform or overpreform their fWAR. This is shown in the next section.

Analysis 2: Several teams consistently under/overperform their predicted fWAR

We now show that several teams consistently win more games or less games than what fWAR predicts. To do this, we compute the average residual across seasons for each team and then divide that average residual value by the corresponding standard error. We will denote these computed quantities as normal scores. Observe that there are 11 franchises out of 38 total franchises with an absolute normal score greater than or equal to 2. There are technically 39 franchises but the Pilots were dropped from consideration because they only played 1 season, and you cannot a standard error from 1 data point.

```
## add intra-season residuals to dataset
WinsvfWAR <- WinsvfWAR %>%
  mutate(resid = do.call(rbind, mclapply(models, mc.cores = 12,
    function(m){
      matrix(m$residuals, ncol = 1)
})))

normalscore <- as.data.frame(WinsvfWAR %>% group_by(Team) %>%
  summarise(normalized = mean(resid)/(sd(resid)/sqrt(length(year)))) %>%
  arrange(desc(abs(normalized))))

## number of franchises (Seattle Pilots dropped)
nrow(normalscore) - 1
```

```
## [1] 38
```

```
## franchises with normal scores >= 2
normalscore %>% filter(abs(normalized) >= 2)
```

```
##            Team normalized
## 1        Giants   3.182230
## 2      Phillies  -3.069237
## 3       Red Sox  -3.008894
## 4        Tigers  -2.961556
## 5      Mariners  -2.889537
## 6     Devil Rays -2.693505
## 7      Athletics  2.444990
## 8          Bees   2.374846
## 9        Braves   2.189987
## 10      Marlins  -2.189601
## 11       Angels   2.007475
```

Under the assumption that there is no relationship between MLB teams and the regression residuals, we expect roughly 2 teams to have an absolute normal score greater than or equal to 2 (corresponding to a test with error tolerance equal to 0.05). However, we observed 11 such teams. The probability of observing 11 or more such teams is computed below.

```
pbinom(10, size = 39, p = 0.05, lower = FALSE)
```

```
## [1] 2.214294e-06
```

This probability is very low, and we would therefore conclude that the assumption of no relationship between regression residuals and MLB teams is violated. In other words, a significantly large number of teams consistently win or lose more games than what is expected by fWAR.

Caveats

- A team's total fWAR is observed at the same time of a team's total wins, so fWAR is not a very practical predictor of wins. However, it may be the case that one can predict next season's fWAR with a high degree of accuracy. From this perspective a team's fWAR can be thought of as an oracle estimator obtained from a projection system and this would make it a practical predictor for wins.

- Three linear modeling assumptions (linear relationship, constant variance, and normality or errors) were verified. However, I did not test for independence of cases. Furthermore, independence of cases does not hold in this context, because a team win comes at a cost of a win to another team. However, the same zero-sum feature holds for the predictor fWAR, an increase of one team's fWAR comes at a cost to the fWAR of another team. I conjecture that the zero-sum nature of team wins is (partially) accounted for by the zero-sum nature of fWAR. For example, I can prove that mean zero regression errors are a condition that guarantees unbiased intercept and slope estimation in this context. A symmetric error structure (with dependence) can be added to facilitate variability estimation. I think that ignoring such structure (which was done in this analysis) produces conservative hypothesis tests for regression coefficient estimates, provided that residuals are constant (which was verified).

- The testing procedure in Analysis 2 that led to 11 franchises having normalized values greater than or equal to 2 requires that the central limit theorem has kicked in so that $\bar{x}/se(\bar{x})$ is approximately normally distributed. This is an assumption that we are reliant on.

- All hypothesis tests were done with error tolerance $\alpha = 0.05$ and tests were not adjusted for multiple comparisons.

Special thanks to CRAN, Rstudio, and Fangraphs!

Thank you for reading!