Factors that Affect Winning Percentage in Major League Baseball

Ryan Steffe

December 8, 2022

**Abstract**

The reputation of Major League Baseball can be summarized with the word consistency. Often you might hear people say, “Stuck-in-their-ways” to describe the game. However, this static image was introduced to change with the Oakland Athletics concept called, “Moneyball”. Moneyball was the coined term to describe the Oakland A’s strategy to utilize data analytics to make decisions (Moneyball). Maybe portrayed as an anomaly at the time, Moneyball paved the way to Major League Baseball’s current reliance on data analysis.

Having said that, this project applied data analysis **to identify the main determinants of winning percentage in Major League Baseball**. Specifically, this project centers around the on-the-field factors (e.g., Batting Average) affecting winning percentage rather than the off-the-field factors. A multiple linear regression model was used to pinpoint these variables. Particularly, the response variable was winning percentage and one independent variable was pitcher strikeouts for example. Data was collected from each team (30 MLB Teams) from the 2010 to 2019 seasons. Altogether, 300 observations were collected to find the answer.

**Previous Research**

Two cases conducted similar research to this topic. Christopher Duquette and his team conducted statistical analysis to determine if the variables relied upon in “Moneyball” were credible. In other words, Duquette wanted to find if variables affected win loss percentage. A least squares regression method was used. The study found that offensive statistics such as singles and homeruns were statistically significant (Duquette, et al). Simply put, singles and homeruns affect win loss percentage.

Another study was carried out by Braden Murray. Murray attempted to discover if baseball metrics excluding runs affect winning percentage. R-squared (R2) values were relied upon to understand the variables influencing winning percentage. Murray discovered that pitching metrics are a more important factor to winning percentage than batting metrics (Murray).

**Methodology**

The question we are attempting to answer in this experiment is:

**What are the main determinants of winning percentage in Major League Baseball?**

Panel data was employed to find the answer. Panel data considers data from specific individuals (i) over time (t) (Condliffe). To explain, data over the course of ten seasons was collected (2010-2019) and would represent “t”. 30 teams were observed per season and would represent “i”. For example, a model could be:

WinPct100it = (Batting Averageit, Walksit, WalksAllowedit).

Furthermore, multiple linear regression was utilized for this experiment. Seven equations were built for the panel data model. The response variable for all the models was winning percentage (WinningPct100). The independent variables varied in each model. The seven equations are listed in figure 1.

|  |
| --- |
| **WinningPct100** = BatAge+PitchersAge+battingaverage100+TotalBases+HitsAllowed+PitcherSO+WalksAllowed |
| **WinningPct100** = BatAge+PitchersAge+PitcherSO+WalksAllowed+Hits+Walks+SacrificeHits+HR9 |
| **WinningPct100** = StrikeOuts+Hits+TotalBases+PitcherSO+H9+BB9 |
| **WinningPct100** = BatAge+PitchersAge+loghits+Walks+PitcherSO+logwalksallowed |
| **WinningPct100** = BatAge+PitchersAge+loghits+logwalks+logStrikeouts+logwalksallowed+PitcherSO |
| **WinningPct100** = BatAge+PitchersAge+SLG100+logwalks+logStrikeouts+logwalksallowed+logPitcherSO |
| **WinningPct100** = battingaverage100+logwalks+logStrikeouts+logHitsAllowed+logwalksallowed+logPitcherSO |

**Figure 1 (Table of Equations)**

The response variable (WinningPct100) is not valued as a decimal but rather as a number. In other words a winning percentage of 50% would be represented as 50.00. BattingAverage100 and SLG100 (Slugging Percentage) is represented similarly. BatAge constitutes as the average age of all batters from each observation (e.g., 27.4). PitchersAge constitutes as the average age of all pitchers from each observation (e.g., 30.2). The logged variables represent a certain variable that logs the total number. To put another way, the variable (Walks) is the total number of walks over the course of a season, logged. Figure 2 displays the five number summary for the variables.



**Figure 2 (Summary Statistics)**

Also, to ensure that the data is credible multicollinearity was checked. All of the equations did not display multicollinearity as the variance inflation factors were below five. However, it is important to note equation 3 did possess a variance inflation factor around 4. This is important because the results for equation 3 could be not as accurate.

**Results and Conclusion**

Figure 3 and figure 4 displays the coefficients and their p-value. Figure 3 contains one equation that contains data from only the 2019 season (OLS1). Figure 3 also contains three of the seven panel model equations, specifically one, two and three. Figure 4 includes equations four, five, six and seven.

A screenshot of a computer

Description automatically generated with low confidence

**Figure 3 (Regression Results)**

**Table

Description automatically generated**

**Figure 4 (Regression Results)**

The regression indicates that some variables are more important than others. For example, pitchers and batters age is statistically significant sometimes, but not other times. Overall, these coefficients are not statistically significant more often than they are. One variable that does influence winning percentage significantly is battingaverage100 and loghits. Battingaverage100 can be interpreted as (Eq1) for a one unit increase in batting average, we expect winning percentage to increase by 1.04. Loghits is interpreted as (Eq4) for a one percent increase in hits we expect winning percentage to increase by 0.42. Another logarithmic variable that is important is logwalksallowed. Within equation 7, as for a one percent increase in walks allowed, we expect winning percentage to decrease by 0.18. The remaining logarithmic variables can be translated similarly. Ongoing, HR9, is the number of homeruns allowed per nine innings. BB9 is the number of walks allowed per nine innings. H9 is the number of hits allowed per nine innings. All of these variables were statistically significant and an important factor in determining winning percentage. HR9’s interpretation is, for one more homerun allowed per nine innings, we expect winning percentage to decrease by 12.36. One final variable to note is slugging percentage (SLG100). The coefficient was statistically significant and is translated as, for a one unit increase in SLG100, we expect winning percentage to increase by 0.86.

Altogether, the variables that are the most important in determining winning percentage are:

* Battingaverage100
* H9
* HR9
* BB9
* SLG100
* All log variables

These variables are the most significant factors in determining winning percentage due to the size of the coefficients and there statistical significance. To explain, these variables possessed larger coefficients than other variables and were a majority of the time statistically different than zero. For these reasons, they are the most important factors in determining winning percentage and should be considered when constructing a Major League Baseball team.

**Code**

#Baseball Project - Panel Data#

#Installing package plm#

install.packages("plm")

library(plm)

library(readxl)

#Reading in Data#

MLB\_Data <- read\_excel("C:/Users/rstef/OneDrive - West Chester University of PA/ECO 400/MLB Data.xlsx", sheet = "Data")

View(MLB\_Data)

MLB\_Data2019 <- read\_excel("C:/Users/rstef/OneDrive - West Chester University of PA/ECO 400/MLB Data Copy.xlsx", sheet = "Data")

View(MLB\_Data2019)

#Adjusting Response Variable (Multiplying by 100)#

MLB\_Data$WinningPct100 <- MLB\_Data$WinningPct\*100

MLB\_Data2019$WinningPct100 <- MLB\_Data2019$WinningPct\*100

#Variable names#

variable.names(MLB\_Data)

#Creating New Log Variables#

MLB\_Data$loghits <- log(MLB\_Data$Hits)

MLB\_Data$logwalksallowed <- log(MLB\_Data$WalksAllowed)

MLB\_Data$logStrikeouts <- log(MLB\_Data$StrikeOuts)

MLB\_Data$logwalks <- log(MLB\_Data$Walks)

MLB\_Data$battingaverage100 <- MLB\_Data$BA\*100

MLB\_Data2019$battingaverage100 <- MLB\_Data2019$BA\*100

MLB\_Data$SLG100 <- MLB\_Data$SLG\*100

MLB\_Data$logHitsAllowed <- log(MLB\_Data$HitsAllowed)

MLB\_Data$logPitcherSO <- log(MLB\_Data$PitcherSO)

#5 number summary#

summary(MLB\_Data)

summary(MLB\_Data2019)

#OLS Regression of 2019 Season#

OLS1 <- lm(WinningPct100~BatAge+PitchersAge+battingaverage100+TotalBases+HitsAllowed+PitcherSO+WalksAllowed, data=MLB\_Data2019)

summary(OLS1)

#Panel Regression#

Eq1 <- lm(WinningPct100~BatAge+PitchersAge+battingaverage100+TotalBases+HitsAllowed+PitcherSO+WalksAllowed, data=MLB\_Data)

summary(Eq1)

Eq2 <- lm(WinningPct100~BatAge+PitchersAge+PitcherSO+WalksAllowed+Hits+Walks+SacrificeHits+HR9, data=MLB\_Data)

summary(Eq2)

Eq3 <- lm(WinningPct100~StrikeOuts+Hits+TotalBases+PitcherSO+H9+BB9, data=MLB\_Data)

summary(Eq3)

#Panel Regression with non-linear model#

Eq4 <- lm(WinningPct100~BatAge+PitchersAge+loghits+Walks+PitcherSO+logwalksallowed, data=MLB\_Data)

summary(Eq4)

Eq5 <- lm(WinningPct100~BatAge+PitchersAge+loghits+logwalks+logStrikeouts+logwalksallowed+PitcherSO, data=MLB\_Data)

summary(Eq5)

Eq6 <- lm(WinningPct100~BatAge+PitchersAge+SLG100+logwalks+logStrikeouts+logwalksallowed+logPitcherSO, data=MLB\_Data)

summary(Eq6)

Eq7 <- lm(WinningPct100~battingaverage100+logwalks+logStrikeouts+logHitsAllowed+logwalksallowed+logPitcherSO, data=MLB\_Data)

summary(Eq7)

#Checking for Multicollinearity#

install.packages("car")

library(car)

vif(Eq1)

vif(Eq2)

vif(Eq3)

vif(Eq4)

vif(Eq5)

vif(Eq6)

vif(Eq7)

#Adding stargazer package#

install.packages("stargazer")

library(stargazer)

stargazer(OLS1, Eq1, Eq2, Eq3, type="text", digits=4)

stargazer(Eq4, Eq5, Eq6, Eq7, type="text", digits=4)

**Work Cited**

Condliffe, Simon

Duquette, Christopher, et al. “Major League Baseballs.” *EBSCOhost*, Nov. 2019,

https://web.s.ebscohost.com/ehost/pdfviewer/pdfviewer?vid=5&sid=ee0882bc-b668-42f8-83d7-d56a61acbcef%40redis.

Murray, Braden. “MLB Winning Percentage Breakdown: Which Statistics Help Teams

Win More Games?” *Samford University*, Apr. 2022,

<https://www.samford.edu/sports-analytics/fans/2022/MLB-Winning-Percentage-Breakdown-Which-Statistics-Help-Teams-Win-More-Games>.

“2010 Major League Baseball Team Statistics”. Baseball Reference, <https://www.baseball->

reference.com/leagues/majors/2010.shtml.