DATASETS AND STORIES

Punching a Ticket to Cooperstown

Philip A. Yates
Mathematical Sciences, DePaul University, Chicago, IL

ABSTRACT
When exposed to principal components analysis for the first time, students can sometimes miss the primary purpose of the analysis. Often the focus is solely on data reduction and what to do after the dimensions of the data have been reduced is ignored. The datasets discussed here can be used as an in-class example, a homework assignment, or a written project, with a focus in this article as an in-class example. The data give the students an opportunity to perform principal components analysis and follow-up analyses on a real dataset that is not necessarily the easiest to handle.

KEYWORDS
Baseball; Classification; Logistic regression; Principal components analysis

1. Introduction

Cooperstown, a small village in upstate New York, has been synonymous with baseball immortality ever since the founding of the National Baseball Hall of Fame and Museum there in 1939. From the stands at the ballpark, to barstools, dorm rooms, and internet chatrooms, baseball fans have long debated what players should and should not be in Major League Baseball’s Hall of Fame.

The rules for being elected to the Hall of Fame have changed quite a bit since the first class of Ty Cobb, Walter Johnson, Christy Mathewson, Babe Ruth, and Honus Wagner was elected in 1936. In its current format adopted in 2014, the Baseball Writer’s Association of America (BBWAA) consider any player for potential election to the Hall of Fame after that player has been retired for five years. Baseball writers eligible to vote can vote for up to 10 candidates on a single ballot. Any candidate receiving votes on at least 75% of the ballots is elected to the Hall of Fame. If the player fails to reach the 75% threshold, they will be on the ballot the next year assuming that they received votes on at least 5% of the ballots cast, otherwise they can no longer appear on the ballot the following year. Players can be on the ballot for up to 10 years. Players that are not elected to the Hall of Fame in this manner are then moved to an Eras Committee system where they can be evaluated and potentially elected into the Hall of Fame at any time (Baseball Hall of Fame 2018a). These Eras Committees are commonly referred to as the “Veterans Committee”; however, they are committees made up of former managers, umpires, executives, and retired players from the following time periods (Baseball Hall of Fame 2018b):

- Today’s Game—1988 to present
- Modern Baseball—1970 to 1987
- Golden Days—1950 to 1969
- Early Baseball—1871 to 1949.

Since players in this dataset have careers that span from Major League Baseball’s early days to the present, it is important to note how players were elected to the Hall of Fame prior to 2014. The election cycles have not always been consistent. From 1936 to 1939, the elections were held annually with no real set rules in place. From 1940 to 1945, the elections were held every three years, meaning they were held in 1942 and 1945 only. This may be due to baseball players serving in World War II. From 1946 to 1956, elections were once again an annual event. From 1957 to 1965, the elections were held every other year. Since 1966, elections have been an annual occurrence. Typically there are no run-off ballots; however, in 1946 the top 20 vote getters comprised a second and final ballot, if no one was elected, and no one was in 1946’s election. In 1949, there was a run-off vote, with Detroit Tigers’ second baseman Charlie Gehringer being elected to the Hall of Fame. From 1960 to 1968, there were run-off ballots where the top 30 vote getters comprised a second and final ballot, if no one was elected. In 1968, there was a limit implemented of 40 players per ballot. In 1945, voters were supposed to focus more on how the game was played on the field instead of a player’s character off the field. In terms of waiting periods for players retired to be eligible for election, from 1936 to 1945 there was no waiting period. From 1946 to 1953, a player needed to have been retired one year. In 1954, a player needed to be retired for five years prior to being eligible for election. In 1979, the “at least appearing on 5% of ballots” rule was introduced. From 1946 to 1955, a player could not be on the ballot if they have been retired for 25 years. This period was extended to 30 years in 1956. In 1962, it was reduced to 20 years. In 2014, it was reduced to its current 15-year period (Baseball Hall of Fame 2018c).

The data described here will focus on position players, that is, nonpitchers, who have received at least one vote for the Hall of Fame at some point in their careers. The dataset will be...
used as an example of principal components analysis as well as using the principal components in a logistic regression model to predict the probability of a player being enshrined into the Hall of Fame. The principal components analysis will be used to reduce the dimensions of the dataset. Logistic regression will then be performed using the retained principal components as predictors to predict the probability of a hitter being inducted into the Hall of Fame.

Specifically in this manuscript, methods and results will be presented in a way for instructors to use the dataset as an in-class example; however, instructors may wish to use this dataset for a homework assignment or as part of a data analysis project. Along the way, helpful hints, potential pitfalls, and alternative applications will be introduced to provide further guidance for any instructors. It should be noted that this type of analysis would work best for upper-level undergraduate mathematics or statistics students and first-year graduate students taking an applied statistics sequence. For the record, these data were specifically used in a master’s level applied statistics sequence at an institution where the sequence has no prerequisites and students need to have taken one linear algebra course during their academic career to be accepted into the program. This type of course might use a textbook like the one by Ramsey and Schafer (2013), where the topics of principal components and logistic regression are addressed.

Helpful Hint: If using a textbook like Ramsey and Schafer's (2013) The Statistical Sleuth, an instructor will have to supplement the material involving principal components analysis with all of the necessary linear algebra needed for students to really grasp what is going on. While a very useful textbook, Ramsey and Schafer omit all of the underlying linear algebra calculations.

2. The Data and the Story

The data in the file baseballHOF.csv consist of all 602 position players that have received at least one vote for Major League Baseball's Hall of Fame. All of the data were compiled from baseball-reference.com (Baseball Reference 2018). The variables are summarized in Table 1. All of these variables except for JAWSratio are found on the website. These metrics, though not a complete list of what can actually be measured on position players in the game, represent a mix of "traditional" and "modern" baseball statistics. The more traditional statistics are the career counting statistics, like batting average, hits, home runs, and runs batted in. Statistics like OPS have started to gain more traction in the baseball community in recent years, even appearing in nationally televised games and on scoreboards in baseball stadiums. Statistics like WAR are considered more "modern" baseball statistics. These attempts to measure a player’s offensive and defensive contributions to their team’s success.

Helpful Hint: Obviously not all students are baseball fans. When introducing these data to the students in class or as an assignment or project, be sure to open the dataset and walk them through the various variables. Take careful consideration when discussing the variables JAWS, Jpos, and JAWSratio. More details about these three variables are in the following Potential Pitfall.

Potential Pitfall: The variables JAWS and Jpos are meant to be compared with one another. Jpos is a variable that basically attempts to keep track of the player’s position, a categorical variable, via a quantitative variable. For example, Chipper Jones, Brooks Robinson, and Mike Schmidt are three Hall of Fame third basemen in the dataset. They all have a Jpos of 55.7. JAWSratio, JAWS divided by Jpos and multiplied by 100, is

| Variable name | Variable description |
|----------------|----------------------|
| Name | Name of the player—identifies the case in the data. |
| HoF | A "Yes"/"No" indicator on whether or not the player is in the Hall of Fame. |
| Yrs | Number of seasons played in player’s career. |
| WAR | Baseball Reference’s measure of Wins Above Replacement. This is a single number that describes the number of wins the player added to their teams over the course of their career. |
| WAR7 | The sum of the seven best seasons of WAR in the player’s career. It may not be seven seasons in a row. |
| JAWS | Developed by Baseball Prospectus. It contains a combination of career and 7-year peak WAR totals allowing for comparison to average Hall of Fame players by position. |
| JAWSratio | The average JAWS score for all Hall of Fame players at this position plus overall Hall of Fame averages for positions with fewer inducted players. |
| Jpos | games played during a player's career. |
| AB | At bats during a player’s career. |
| R | Runs scored during a player’s career. |
| H | Hits during a player’s career. |
| HR | Home runs hit during a player’s career. |
| RBI | Runs batted in during a player’s career. |
| SB | Stolen bases during a player’s career. |
| BB | Walks (bases on balls) during a player's career. |
| BA | Batting average. This is H divided by AB. |
| OBP | On base percentage. This is the sum of H, BB, and times hit by a pitch divided by the sum of AB, BB, times hit by a pitch, and sacrifice flies. |
| SLG | Slugging percentage. This is the number of bases divided by AB. Every single is one base, double is two bases, triple is three bases, and HR is four bases in the numerator of this calculation. |
| OPS | The sum of OBP and SLG. |
| OPSadj | OPS adjusted to the player's ball park. 100 is an "average" hitter. |
### Table 2. Summary statistics.

|                | Not in Hall of Fame |               | Hall of Fame |               |
|----------------|---------------------|---------------|--------------|---------------|
|                | Mean                | SD            | Min.         | Q1            | Median         | Q3            | Max.         |
| Yrs            | 14.96               | 3.37          | 4.00         | 13.00         | 15.00          | 17.00         | 24.00        |
| WAR            | 30.58               | 17.67         | -5.30        | 17.60         | 29.10          | 40.80         | 162.80       |
| WAR7           | 24.12               | 10.35         | 0.00         | 16.75         | 24.40          | 30.95         | 72.20        |
| JAWSratio      | 50.239              | 25.05         | -4.55        | 31.24         | 50.35          | 65.19         | 220.19       |
| G              | 1662                | 486.11        | 67           | 1337          | 1673           | 1996          | 3592         |
| AB             | 5836                | 1872.89       | 186          | 4588          | 5923           | 7129          | 14053        |
| R              | 834.6               | 325.17        | 25.0         | 617.5         | 832.0          | 1031.5        | 2227.0       |
| H              | 1622                | 549.82        | 54           | 1252          | 1621           | 1992          | 4256         |
| HR             | 145.5               | 127.47        | 1.0          | 39.0          | 105.0          | 231.5         | 762.0        |
| RBI            | 776.7               | 344.26        | 24.0         | 513.5         | 728.0          | 1026.5        | 1996.0       |
| BB             | 125.7               | 129.99        | 2.0          | 37.5          | 80.0           | 172.0         | 752.0        |
| BB             | 600.2               | 306.18        | 7.0          | 382.5         | 559.0          | 792.5         | 2558.0       |
| BA             | 0.277               | 0.020         | 0.208        | 0.264         | 0.276          | 0.290         | 0.356        |
| OBP            | 0.345               | 0.029         | 0.254        | 0.326         | 0.345          | 0.364         | 0.444        |
| SLG            | 0.413               | 0.060         | 0.256        | 0.369         | 0.411          | 0.455         | 0.607        |
| OPS            | 0.758               | 0.081         | 0.529        | 0.704         | 0.760          | 0.811         | 1.051        |
|OPSadj          | 107.1               | 19.85         | 22.0         | 94.0          | 109.0          | 120.0         | 182.0        |

calculated so that any ratio greater than 100 means that the player had a higher value of JAWS compared to all other Hall of Fame players at that given position while a ratio less than 100 means that the player had a lower value of JAWS compared to all other Hall of fame players at that given position. When incorporating this information into the model, encourage students to use JAWSratio since it captures information from JAWS and Jpos.

### 3. The Analysis

All analyses were conducted using R version 3.4.3 (R Core Team 2018). Table 2 contains summary statistics for the 17 quantitative variables of interest for position players in the Hall of Fame and those position players who received votes but failed to make the Hall of Fame. Table 3 is the correlation matrix for the 17 quantitative variables of interest. If one includes all of these 17 variables as predictors in a model, severe multicollinearity occurs. The results of said model can be found in Table 4. It should be noted that the variance inflation factors for a logistic regression model using all of these variables as predictors range from 2.786 for stolen bases (SB), the only one less than 5, to 3217.564 for OPS. These variance inflation factors were calculated using the car package in R (Fox and Weinberg 2011). For more details, see the code in the online appendix. Since the variables are all on different scales, principal components based off the correlation matrix is the most appropriate method here. Using principal components analysis can help solve the problem of multicollinearity in the model. In practice, multicollinearity does not cause any issues with the model’s ability to make predictions; however, the standard errors of the regression coefficients tend to be inflated in the presence of other highly correlated predictor variables (Neter et al. 1996). If a researcher is using a variable selection method like forward, backward, or stepwise selection based on the t-statistic, for example, potentially important predictors will be discarded from the model since their possible larger standard errors will make the predictor seem less significant than what it might actually be.

Potential Pitfall: If not given proper guidance on an assignment or project, students may just include all of the predictors into a logistic regression model and ignore any issues pertaining to multicollinearity since it generally will not impact the model’s
Table 3. Correlation matrix for the 17 quantitative variables of interest (lower triangle).

|        | Yrs | WAR | WAR7 | JAWS | G   | AB  | R   | H   | HR  |
|--------|-----|-----|------|------|-----|-----|-----|-----|-----|
| Yrs    | 1.000 |     |      |      |     |     |     |     |     |
| WAR    | 0.595 | 1.000 |      |      |     |     |     |     |     |
| WAR7   | 0.504 | 0.954 | 1.000 |      |     |     |     |     |     |
| JAWS   | 0.580 | 0.968 | 1.000 | 1.000 |     |     |     |     |     |
| G      | 0.847 | 0.733 | 0.679 | 0.712 | 1.000 |     |     |     |     |
| AB     | 0.779 | 0.734 | 0.691 | 0.710 | 0.973 | 1.000 |     |     |     |
| R      | 0.675 | 0.849 | 0.804 | 0.814 | 0.849 | 0.896 | 1.000 |     |     |
| H      | 0.743 | 0.786 | 0.742 | 0.759 | 0.927 | 0.972 | 0.927 | 1.000 |     |
| HR     | 0.409 | 0.567 | 0.574 | 0.558 | 0.512 | 0.541 | 0.478 | 0.811 | 1.000 |
| RBI    | 0.651 | 0.773 | 0.750 | 0.764 | 0.792 | 0.796 | 0.811 | 0.818 | 0.807 |
| SB     | 0.303 | 0.368 | 0.320 | 0.328 | 0.346 | 0.409 | 0.423 | −0.126 | 0.117 |
| BB     | 0.589 | 0.794 | 0.751 | 0.773 | 0.738 | 0.688 | 0.793 | 0.672 | 0.635 |
| BA     | 0.143 | 0.487 | 0.508 | 0.482 | 0.216 | 0.307 | 0.496 | 0.507 | 0.106 |
| BA     | 0.143 | 0.487 | 0.508 | 0.482 | 0.216 | 0.307 | 0.496 | 0.507 | 0.106 |
| OBP    | 0.204 | 0.637 | 0.663 | 0.640 | 0.282 | 0.288 | 0.568 | 0.425 | 0.328 |
| SLG    | 0.224 | 0.576 | 0.633 | 0.593 | 0.345 | 0.351 | 0.520 | 0.439 | 0.351 |
| SLG    | 0.224 | 0.576 | 0.633 | 0.593 | 0.345 | 0.351 | 0.520 | 0.439 | 0.351 |
| OPS    | 0.237 | 0.647 | 0.698 | 0.660 | 0.352 | 0.359 | 0.582 | 0.472 | 0.472 |
| OPS    | 0.237 | 0.647 | 0.698 | 0.660 | 0.352 | 0.359 | 0.582 | 0.472 | 0.472 |
| OPSadj | 0.273 | 0.704 | 0.744 | 0.713 | 0.367 | 0.371 | 0.588 | 0.480 | 0.624 |
| OPSadj | 0.273 | 0.704 | 0.744 | 0.713 | 0.367 | 0.371 | 0.588 | 0.480 | 0.624 |

The goal of principal components analysis is to transform the original variables into a system of orthogonal variables and to use as few of these transformed variables as possible to describe the variation in the original dataset. Each hitter’s offensive variables are transformed via

\[ z_i = Ay_i^*, \quad i = 1, \ldots, 602, \]

where \( z_i \) represents the principal component scores for the ith hitter, \( A \) is an orthogonal matrix, and \( y_i^* \) are the standardized variables for the ith hitter. The correlation matrix of the offensive variables, \( R \), can be found in Table 3. The sample covariance matrix of \( z \) will be a diagonal matrix with \( s_{z_{11}}, \ldots, s_{z_{17}} \) on the diagonal. This matrix will be equal to \( A R A^T \). Rencher (1995) employed a spectral decomposition argument to state that since \( A \) is an orthogonal matrix, the values \( s_{z_{p}}, \quad p = 1, \ldots, 17 \), will be equal to the eigenvalues of \( R \). Also, the rows of \( A \) are the normalized eigenvectors of \( R \). The eigenvalues for the correlation matrix can be found in Table 5. These will be investigated to determine the number of principal components that will be necessary to use in this analysis.

Rencher (1995) proposed some guidelines to assist in this process. Two of the simplest for students to understand is retaining enough components so that at least 80% of the total variance is explained and ignoring components whose eigenvalues are less than one. This latter criterion is known as the Kaiser criterion or Kaiser–Guttman rule (Guttman 1954; Kaiser 1970). Basically, if the measured variables are standardized, they have a variance equal to one. The eigenvalues represent the variance of the principal component. If the eigenvalue is less than one, it fails to account for as much variability as a single measured variable.

Table 4. Estimated coefficients for the logistic regression model: all predictors.

|        | Estimate | SE   | z-value | p-value | VIF |
|--------|----------|------|---------|---------|-----|
| Intercept | −29.521 | 10.324 | 2.859 | 0.004 |     |
| Yrs | 0.196 | 0.107 | 1.827 | 0.068 | 5.265 |
| WAR | −0.045 | 0.044 | −1.021 | 0.307 | 24.771 |
| WAR7 | 0.092 | 0.053 | 1.732 | 0.083 | 8.949 |
| JAWS | 0.075 | 0.029 | 2.628 | 0.009 | 18.262 |
| G | −0.004 | 0.003 | −1.294 | 0.196 | 70.577 |
| AB | 0.002 | 0.002 | 1.378 | 0.168 | 305.547 |
| R | 0.002 | 0.002 | 1.157 | 0.247 | 16.338 |
| HR | −0.005 | 0.004 | −1.148 | 0.251 | 217.261 |
| RBI | 0.004 | 0.001 | 3.022 | 0.004 | 9.329 |
| SB | −0.001 | 0.001 | −0.537 | 0.591 | 2.786 |
| BB | −0.005 | 0.002 | −2.229 | 0.026 | 23.187 |
| BA | 30.070 | 37.774 | 0.796 | 0.426 | 24.439 |
| OBP | 124.293 | 115.247 | 1.078 | 0.281 | 391.423 |
| SLG | 60.916 | 112.795 | 0.540 | 0.589 | 2083.015 |
| OPS | −71.714 | 112.334 | −0.638 | 0.523 | 3217.564 |
| OPSadj | −0.031 | 0.022 | −1.385 | 0.166 | 6.715 |

fit or its ability to make predictions (Neter et al. 1996). While performing principal components analysis does not address multicollinearity in a linear model, it can be useful in this situation since principal components are linearly uncorrelated with one another.

fit or its ability to make predictions (Neter et al. 1996). While performing principal components analysis does not address multicollinearity in a linear model, it can be useful in this situation since principal components are linearly uncorrelated with one another.
However, Preacher and MacCallum (2003) said researchers misapply this rule because they ignore the fact that it was originally supposed to be a lower bound on the number of components in principal components analysis or factors in exploratory factor analysis to retain. Since the goal of this analysis is to use principal components as predictors in a logistic regression model, the approach taken will be one similar to one taken by Milewska et al. (2014). A variable selection process, specifically a backward selection procedure using p-values from the z-tests for the estimated logistic regression coefficients for this baseball analysis, will be used to determine how many principal components will be retained in the model. Using this method, potentially significant linear combinations of the hitters’ variables will not be ignored, even if the components do not necessarily explain a large proportion of variation of the original variables. The results of this variable selection procedure can be found in Table 6. The original dataset of 17 variables has now been reduced to eight principal components.

Table 7 shows the eigenvectors for the eight retained components of the quantitative variables of interest. These eigenvectors represent the coefficients for each of the eight retained principal components. Denoting the principal components as $z_i$, as seen previously, the principal components are

$$z_i = a_i^T (S^{1/2})^{-1} (y - \bar{y}) = \sum_{j=1}^{17} a_{ij} \left( \frac{y_j - \bar{y}_j}{s_{y_j}} \right),$$

$i = 1, 3, 4, 5, 6, 11, 14, 16$.

### 3.1. Interpreting the Principal Components

There is some debate on how to best interpret principal components. Preacher and MacCallum (2003) stated because principal components analysis does not explicitly model error variance, it causes problems when attempting any interpretations of substance with respect to the principal components. They say that principal components analysis should only be used when the goal is to find linear combinations of model variables that retain as much of the variance in the model variables as possible or find components that explain as much variance as possible. If the goal is to find interpretable constructs that explain correlations among the model variables, exploratory factor analysis should be the analysis of choice.

Sharma (1996) argued that because the principal components are linear combinations of variables of some interest, it is only natural to attempt to assign some meaning to the components. He stated that one should focus on the loadings of each variable to determine which variables are influential in the formation of the component. Although no additional citations were mentioned, Sharma suggested that a loading of 0.5 or above, in magnitude, is a good rule of thumb to determine which variables are influential to the component.

Rencher (1995) discussed how certain patterns in the loadings and the correlation matrix can often aid in the interpretation of the principal components. If all of the loadings or elements of the first eigenvector are positive, this is sometimes referred to as a measure of size. This is analogous to the first principal component being a weighted average of the variables that make up the component. Correlations between the variables and the principal components can be computed to aid in the interpretation of the components. To calculate the correlation between the $i$th principal component and the $j$th standardized variable, take the coefficient associated with the $j$th standardized variable in the $i$th principal component and multiply that by the square root of the $i$th eigenvalue of the correlation matrix (see Table 5 for the eigenvalues). Table 8 lists the correlations between the retained principal components and the quantitative variables of interest.

Based on the variables’ coefficients or loadings in each of the retained principal components (Sharma’s proposed technique) as well as the corresponding correlations between each standardized variable and principal component (Rencher’s proposed technique), possible interpretations are as follows:

- $z_1$ represents a weighted average of the standardized variables since all of the loadings are positive,
- $z_2$ represents a difference in speed versus power in terms of SB and home runs, due to the large magnitudes of the loadings for those two variables,
- $z_4$ represents the batting average, or possibly the difference of hitting to get on base versus getting on base via other methods, that is, bases on balls,
- $z_5$ represents the negative of the standardized value of SB, perhaps penalizing players with poor base running skills.

Each of these previously mentioned principal components either have strong correlations with the specified standardized variables or have loadings or coefficients at least 0.5 in magnitude for each of the mentioned standardized variables. Scatterplots of the interpreted principal components versus the principal component scores can be found in Figure 1.

Alternative Application: For the last four retained principal components, the interpretation is not as obvious mainly due to the small magnitudes for the loadings or coefficients and the very weak correlations with the corresponding standardized variables. That makes these interpretations not as obvious. After walking through the thought process on the previous interpretations, ask students to attempt possible interpretations of the last four retained principal components.
Principal components to overcome the issue of multicollinearity in this project (or exercise or example), students created principal components. For this project (or exercise or example), students created principal components. Correlation matrix for the 17 quantitative variables of interest and retained principal components.

### Table 6. Results of backward selection procedure: z-statistic p-values with $\alpha = 0.05$

| Variable | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ | $a_6$ | $a_{11}$ | $a_{14}$ | $a_{16}$ |
|----------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| Yrs      | 0.203 | -0.128| 0.141 | 0.254 | -0.268| -0.101| -0.202| -0.024|       |
| WAR      | 0.284 | 0.094 | -0.258| 0.131 | 0.339 | 0.281 | 0.595 | -0.155|       |
| WAR7     | 0.279 | 0.091 | -0.259| 0.096 | 0.442 | -0.419| 0.165 | 0.056 |       |
| JAWSratio| 0.281 | 0.078 | -0.272| 0.164 | 0.397 | 0.099 | -0.685| 0.061 |       |
| G        | 0.251 | -0.135| 0.114 | 0.089 | -0.084| 0.149 | 0.274 | 0.141 |       |
| AB       | 0.252 | -0.053| 0.212 | -0.005| 0.013 | -0.061| -0.174| -0.711|       |
| R        | 0.282 | 0.109 | 0.053 | -0.129| -0.109| -0.424| 0.008 | 0.030 |       |
| H        | 0.268 | 0.059 | 0.308 | 0.046 | 0.047 | 0.072 | -0.060| 0.652 |       |
| HR       | 0.215 | -0.512| -0.090| -0.355| -0.002| 0.391 | 0.006 | 0.006 |       |
| RBI      | 0.278 | -0.243| 0.212 | -0.112| 0.059 | -0.089| 0.014 | -0.017|       |
| SB       | 0.098 | 0.536 | -0.166| -0.703 | -0.121| 0.091 | -0.006| 0.006 |       |
| BB       | 0.257 | -0.075| -0.443| 0.157 | -0.472| 0.221 | -0.155| 0.015 |       |
| BA       | 0.174 | 0.432 | 0.531 | 0.160 | 0.075 | 0.401 | -0.046| -0.124|       |
| OBP      | 0.215 | 0.293 | -0.148| 0.279 | -0.433| -0.112| 0.098 | -0.027|       |
| SLG      | 0.226 | -0.188| -0.168| -0.273| 0.016 | -0.228| -0.014| 0.001 |       |
| OPS      | 0.241 | -0.032| -0.071| -0.100| -0.144| -0.210| 0.037 | 0.007 |       |
| OPSadj   | 0.242 | 0.040 | -0.009| -0.104| -0.030| 0.163 | -0.051| 0.004 |       |

**NOTE:** A bold value represents the p-value from the z-test for the estimated logistic regression coefficient of the variable eliminated at that specific iteration of the backward selection procedure using a significance level of 0.05.

### Table 7. Eigenvectors for the correlation matrix for the retained components.

| Variable | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ | $a_6$ | $a_{11}$ | $a_{14}$ | $a_{16}$ |
|----------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| Yrs      | 0.664 | -0.155| 0.121 | 0.177 | -0.176| -0.022| -0.002| -0.001|       |
| WAR      | 0.926 | 0.113 | -0.220| 0.091 | 0.222 | 0.061 | 0.062 | -0.008|       |
| WAR7     | 0.912 | 0.109 | -0.221| 0.067 | 0.277 | -0.091| 0.017 | 0.003 |       |
| JAWSratio| 0.917 | 0.094 | -0.233| 0.114 | 0.261 | 0.022 | -0.071| 0.003 |       |
| G        | 0.817 | -0.163| 0.097 | 0.062 | -0.055| 0.032 | 0.029 | 0.007 |       |
| AB       | 0.824 | -0.064| 0.181 | -0.003| 0.009 | -0.013| -0.018| -0.035|       |
| R        | 0.921 | 0.132 | 0.045 | -0.090| -0.071| -0.092| 0.001 | 0.002 |       |
| H        | 0.875 | 0.071 | 0.264 | 0.032 | 0.031 | 0.016 | -0.006| 0.032 |       |
| HR       | 0.703 | -0.618| -0.077| -0.247| -0.001| 0.085 | 0.001 | 0.000 |       |
| RBI      | 0.906 | -0.293| 0.181 | -0.077| 0.039 | -0.019| 0.001 | -0.001|       |
| SB       | 0.319 | 0.646 | -0.142| -0.488| -0.080| 0.020 | -0.001| 0.000 |       |
| BB       | 0.837 | -0.091| -0.379| 0.109 | -0.310| 0.048 | -0.016| 0.001 |       |
| BA       | 0.568 | 0.521 | 0.454 | 0.111 | 0.050 | 0.087 | -0.005| -0.006|       |
| OBP      | 0.700 | 0.353 | -0.126| 0.194 | -0.285| -0.024| 0.010 | -0.001|       |
| SLG      | 0.736 | -0.226| 0.144 | -0.189| 0.011 | -0.050| -0.001| 0.000 |       |
| OPS      | 0.787 | -0.038| 0.061 | -0.069| -0.095| -0.046| -0.004| 0.000 |       |
| OPSadj   | 0.792 | 0.048 | -0.042| -0.072| -0.020| 0.035 | -0.005| 0.000 |       |

### 3.2. Assessing the Model Fit

For this project (or exercise or example), students created principal components to overcome the issue of multicollinearity in a logistic regression model. A logistic regression model was fit using whether or not the player is in the Hall of Fame as a response variable, where a “success” is that the player is in the Hall of Fame, and the eight principal components as the predictor variables. Table 9 contains the estimated coefficients for the logistic regression model with the eight principal components.
as the predictor variables. The most significant predictor is the first retained principal component, $z_1$, given all of the other predictors in the model. Recalling that this principal component represents the weighted average of the standardized original variables, it is not surprising that the coefficient is positive, meaning that as a hitter’s metrics become more large than average, it is more likely that the hitter will be elected to the Hall of Fame, given all other predictors are held constant. The second
retained principal component, \( z_3 \), is the second most significant predictor, given all of the other predictors in the model. This principal component was interpreted as being a measure of speed versus power based on a difference in the standardized value of SB and the standardized value of home runs. Since this estimated regression coefficient is positive, the larger the difference in these standardized values, the more likely the hitter will be inducted into the Hall of Fame, given that all other predictors are held constant. This difference becomes more positive the more SB a hitter has than average compared to the more home runs a hitter has than average. The principal component \( z_3 \) was interpreted to be the difference in the standardized value of batting average and the standardized value of bases on balls. Since this estimated regression coefficient is positive, as a hitter’s batting average becomes better than average compared to a hitter’s bases on balls compared to the average, they are more likely to be inducted into the Hall of Fame, given that all other predictors are held constant. The principal component \( z_5 \) was interpreted to be basically the negative of the standardized value of SB. Since this estimated regression coefficient is positive, larger values of \( z_5 \) mean the hitter is more likely to be inducted into the Hall of Fame, given that all other predictors are held constant. All things being equal, if a hitter has fewer SB than average, they are more likely to be inducted into the Hall of Fame. This may sound like it is basically the opposite conclusion of the year rem like yel to be inducted into the Hall of Fame. The following discussion about Tables 8 and 9 will help the students see where the models went “wrong.”

Table 10 shows the predicted classification of each player based on the results from the logistic regression model. Overall, the model correctly classifies 89.7% of the players (540 out of 602 players). Out of the 447 players not in the Hall of Fame, the model correctly identified 95.1% of them (425 out of 447). For the 155 players in the Hall of Fame, the model correctly identified 74.2% of them (115 out of 155).

Helpful Hint: Encourage students to investigate the actual cases that are false negatives (actual Hall of Fame players identified as not worthy of selection) and false positives (players not elected to the Hall of Fame that model thinks should be in the Hall of Fame). The following discussion about Tables 8 and 9 will help the students see where the models went “wrong.”

Table 11 shows the “false positives” in the model, that is, the players predicted to be enshrined into the Hall of Fame but who are not actually members of the Hall of Fame. Of the twenty-three players on the list, five of them (Barry Bonds, Larry Walker, Scott Rolen, Edgar Martinez, Manny Ramirez) are all still eligible to be elected by the BBWAA; however, two of the five players mentioned (Bonds and Ramirez) might not ever be elected to the Hall of Fame due to their ties, either confirmed or alleged, to performance enhancing drugs (Fainaru-Wada and Williams 2006; Tayler 2014). Pete Rose, baseball’s all-time leader in hits, and Shoeless Joe Jackson are two players that are permanently banned from the Hall of Fame for having gambled on baseball games. Bill Dahlen was considered by many to be one of the greatest shortstops in baseball during his playing career from 1891 to 1911; however, “Bad Bill” supposedly fought with teammates, managers, and umpires and anecdotal had a bit of a drinking problem (Womack 2015). These “character issues” may have long turned off voters from voting for him. Many believe that voters have used a similar rationale of character issues for not inducting Dick Allen into the Hall of Fame, even though sabermetricians like Bill James once called Allen “the most gifted baseball player I have ever seen” (James 1987). The player in this table that might be the most curious to see for baseball fans is Nomar Garciaparra. He won the American League Rookie of the Year Award in 1997, came in second in voting for the American League Most Valuable Player Award in 1998, and was selected six times to Major League Baseball’s All Star Game; however, his career was cut short due to injuries.

Table 12 shows the “false negatives” predicted by the model, that is, the players predicted not to be in the Hall of Fame but who are actually members of the Hall of Fame. Twenty-nine out of the forty players (72.5%) were elected by the veterans committees or one of the eras committees. Most of these players elected by these committees, denoted with a dagger in Table 12, played in the pre-integration era of Major League Baseball (prior to 1947, Jackie Robinson’s debut in Major League Baseball). Some of these players played in what is known as the “dead ball” era of baseball, roughly the years 1900 to 1919, prior to Babe Ruth’s emergence as a hitter rather than a pitcher. The game of baseball during this era was more strategy based. Managers would rely on SB and hit-and-run plays—which are plays in which the base runner starts running as the pitcher throws to home plate creating an opportunity to take more bases with a base hit, than on simply using home runs to score (Okrent, Lewine, and Nemec 2000). Players on this list from the dead
Figure 2. Deviance residual plots versus fitted values and retained principal components.

Table 10. Player classification: logistic regression model results.

| Actual | Model prediction | No HoF | HoF |
|--------|------------------|-------|-----|
| No HoF | 425              | 22    |
| HoF    | 40               | 115   |

Table 11. Player classification: predicted Hall of Famers not in the Hall of Fame.

| Name               | $\hat{P}$ (HoF) | Name               | $\hat{P}$ (HoF) |
|--------------------|-----------------|--------------------|-----------------|
| Pete Rose          | 0.990           | Edgar Martinez     | 0.671           |
| Barry Bonds        | 0.986           | Vada Pinson        | 0.660           |
| Shoeless Joe Jackson| 0.982          | Buddy Bell         | 0.660           |
| Bill Dahlen        | 0.888           | Ken Boyer          | 0.624           |
| Lave Cross         | 0.800           | Sherry Magee       | 0.611           |
| Nomar Garciaparra  | 0.765           | Minnie Minoso      | 0.590           |
| Kenny Lofton       | 0.740           | Keith Hernandez    | 0.582           |
| Larry Walker       | 0.709           | Manny Ramirez      | 0.571           |
| Bobby Veach        | 0.688           | Dick Allen         | 0.531           |
| Scott Rolen        | 0.688           | Thurman Munson     | 0.515           |
| Ted Simmons        | 0.678           | Ken Williams       | 0.508           |

ball era may have had statistics greater than their peers; however, their statistics pale in comparison to even above average players in its current iteration.

4. Conclusion

The previously discussed analysis is an example on how to use principal components beyond simply reducing the dimensions of a dataset. While one would hope that the logistic regression model’s ability to correctly classify players would be better than what is presented here, it gives students another example of the
“joys” of working with real data. After all, the cleanest examples appear in textbooks’ examples for a reason.

The plan is to keep adding to this dataset with every annual vote for hitters’ induction into the Hall of Fame, rerunning the model, and predict which hitters would and would not be in the Hall of Fame. Perhaps with a newer generation of more statistical savvy baseball writers and Hall of Fame voters, the number of “false positives” will decrease and the number of “false negatives” will not increase, making it pretty obvious which hitters are worthy of induction. As far as the “false negatives” go, there is nothing that can be done to really reduce those numbers. Once a hitter is inducted, they cannot be retroactively removed from the Hall of Fame.

Students may wish to further explore principal components analysis, logistic regression, and baseball’s Hall of Fame. This analysis only focuses on hitters. A natural follow-up analysis or project could entail running a similar type of analysis on the National Hockey League, just to name a few professional sports leagues in North America.

Instructors wishing to use this dataset in their classrooms are not just limited to using principal components analysis or logistic regression. These data can be used in introductory statistics classes when discussing the robustness of the two-sample t-test or performing a randomization test on a variable of one’s choice to compare the difference between Hall of Fame and non-Hall of Fame players. Even though not all students follow

### Table 12. Player classification: Hall of Famers predicted not to be in the Hall of Fame.

| Name               | Years               | \( \hat{P} \text{ (HoF)} \) | Name               | Years               | \( \hat{P} \text{ (HoF)} \) |
|--------------------|---------------------|-----------------------------|--------------------|---------------------|-----------------------------|
| Red Schoendienst†  | 1945–1963           | 0.490                       | Jim Thome*         | 1991–2012           | 0.226                       |
| Gabby Hartnett*    | 1922–1941           | 0.467                       | Ross Youngs†       | 1917–1926           | 0.225                       |
| Dave Winfield†     | 1973–1988           | 0.463                       | Ralph Kiner*       | 1946–1955           | 0.225                       |
| Bobby Doerr†       | 1937–1944, 1946–1951| 0.419                       | Lloyd Waner†       | 1927–1942, 1944–1945| 0.203                       |
| Dave Bancroft†     | 1915–1930           | 0.417                       | Freddie Lindstrom† | 1924–1936           | 0.202                       |
| Jim Bottomley†     | 1922–1937           | 0.410                       | Larry Doby†        | 1947–1959           | 0.200                       |
| Tony Perez*        | 1964–1986           | 0.395                       | George Kell†       | 1943–1957           | 0.194                       |
| Enos Slaughter†    | 1938–1942, 1946–1959| 0.395                       | Ernie Lombardi†    | 1931–1947           | 0.182                       |
| Max Carey†         | 1910–1929           | 0.381                       | Willie McCovey*    | 1959–1980           | 0.174                       |
| Lou Brock*         | 1961–1979           | 0.368                       | Joe Tinker†        | 1902–1916           | 0.132                       |
| Travis Jackson†    | 1922–1936           | 0.349                       | Ray Campanella*    | 1948–1957           | 0.109                       |
| Reggie Jackson*    | 1967–1987           | 0.335                       | Roger Bresnaham†   | 1897, 1900–1915     | 0.107                       |
| Harry Hooper†      | 1909–1925           | 0.305                       | High Pockets Kelly†| 1915–1917           | 0.103                       |
| Joe Gordon†        | 1938–1943, 1946–1950| 0.302                       | Phil Rizzuto†      | 1941–1942, 1946–1956| 0.096                       |
| Tony Lazzeri†      | 1926–1939           | 0.295                       | Johnny Evers†      | 1902–1917, 1922, 1929| 0.092                       |
| Hack Wilson†       | 1923–1934           | 0.290                       | Harmon Killebrew*  | 1954–1975           | 0.069                       |
| Orlando Cepeda†    | 1958–1974           | 0.275                       | Bill Mazeroski†    | 1956–1972           | 0.042                       |
| Frank Chance†     | 1898–1914           | 0.259                       | Rick Ferrell†      | 1929–1945, 1947    | 0.028                       |
| Willie Stargell*   | 1962–1982           | 0.257                       | Tommy McCarthy†    | 1884–1896           | 0.025                       |
| Chick Hafey†       | 1924–1935, 1937     | 0.239                       | Ray Schalk†        | 1912–1929           | 0.008                       |

†Elected by one of the veterans or eras committees.

*Elected by the BBWAA.

**Supplementary Materials**

The R code was used to read the file `baseballHOF.csv` into R, fit the multiple regression model, calculate the variance inflation factors, and to generate the correlation matrix. All figures were generated using the ggplot2 package (Wickham 2016) in R.

**References**

Baseball Hall of Fame (2018a), “BBWAA Election Rules,” available at [https://baseballhall.org/hall-of-famers/rules/bbwaa-rules-for-election](https://baseballhall.org/hall-of-famers/rules/bbwaa-rules-for-election).

——— (2018b), “Eras Committees,” available at [https://baseballhall.org/hall-of-famers/rules/eras-committees](https://baseballhall.org/hall-of-famers/rules/eras-committees).

——— (2018c), “Voting Rules History,” available at [https://baseballhall.org/hall-of-famers/rules/voting-rules-history](https://baseballhall.org/hall-of-famers/rules/voting-rules-history).

Baseball Reference (2018), “Hall of Fame Voting,” available at [https://www.baseball-reference.com/awards/hof_2018.shtml#all_hof_BBWAA](https://baseball-reference.com/awards/hof_2018.shtml#all_hof_BBWAA).

Fainaru-Wada, M., and Williams, L. (2006), Game of Shadows: Barry Bonds, BALCO, and the Steroids Scandal That Rocked Professional Sports, New York: Gotham Books.

Fox, J., and Weisberg, S. (2011), An R Companion to Applied Regression (2nd ed.), Thousand Oaks, CA: SAGE.

Guttman, L. (1954), “Some Necessary Conditions for Common-Factor Analysis,” Psychometrika, 19, 149–161.

James, B. (1987), The Bill James Historical Baseball Abstract, New York: Villard Books.

Kaiser, H. F. (1970), “A Second Generation Little Jiffy,” Psychometrika, 35, 401–415.

Lele, S. R., Keim, J. L., and Solymos, P. (2017), “ResourceSelection: Resource Selection (Probability) Functions for Use-Availability Data,” R
Milewska, A. J., Jankowska, D., Citko, D., Wiesak, T., Acacio, B., and Milewski, R. (2014), “The Use of Principal Component Analysis and Logistic Regression in Prediction of Infertility Treatment Outcome,” *Studies in Logic, Grammar, and Rhetoric*, 39, 7–23, DOI: 10.2478/slgr-2014-0043.

Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. (1996), *Applied Linear Statistical Models* (4th ed.), Chicago, IL: Irwin.

Okrent, D., Lewine, H., and Nemec, D. (2000), *The Ultimate Baseball Book: The Classic Illustrated History of the World's Greatest Game*, Boston, MA: Houghton Mifflin Company.

Preacher, K. J., and MacCallum, R. C. (2003) "Repairing Tom Swift's Electric Factor Analysis Machine," *Understanding Statistics*, 2, 13–43.

Ramsey, F. L., and Schafer, D. W. (2013), *The Statistical Sleuth: A Course in Methods of Data Analysis* (3rd ed.), Boston, MA: Brooks/Cole, Cengage Learning.

Rencher, A. C. (1995), *Methods of Multivariate Analysis*, New York, NY: John Wiley & Sons, Inc.

Sharma, S. (1996), *Applied Multivariate Techniques*, Hoboken, NJ: Wiley.

Tayler, J. (2014), "Manny Ramirez Opens Up on PED Suspensions, Wants to Return to Baseball," *Sports Illustrated*, available at https://www.si.com/mlb/strike-zone/2014/03/13/manny-ramirez-interview-ped-suspension-mlb-return.

Wickham, H. (2016), *ggplot2: Elegant Graphics for Data Analysis*, New York, Springer-Verlag.

Womack, G. (2015), "Why Has Bill Dahlen's Hall of Fame Induction Taken So Long?", *Sporting News*, available at http://www.sportingnews.com/us/mlb/news/bill-dahlen-hall-of-fame-chances-stats-chicago-colts-orphans-brooklyn/n9j4u52dvjcm1oe5mgd96vi.