Teaching Data Analysis with Baseball: A Study of Pitching Strategy in Power Five Conference Teams

Authors
Alan Dabney, Mark Cahill, Drew Kearny, Avery Slack

Keywords
Teaching Statistics; Baseball; Pitching; Linear Regression; Logistic Regression; Multicollinearity; Exploratory Data Analysis

Abstract
This study explores a rich college baseball dataset and demonstrates how the data can be used in the classroom as a case study for teaching data analysis. The dataset comprises team and player statistics for 61 Power Five conference schools during 2021-2023 using Trackman and Sportradar's Synergy Sports technology software. Variables include measures of teams' overall strengths, such as win percentage and playoff participation, as well as performance characteristics of pitchers, such as exit velocity and strikeout percentage. The focus of the dataset is on examining the relationship between pitching strategy and team performance. We present two specific modeling exercises: (1) linear modeling of the rating percentage index (RPI), a widely-used metric for ranking sports teams, and (2) logistic regression modeling of playoff qualification. Along the way, we highlight interesting "stories" in the data suitable for engaging students in learning statistical topics such as exploratory data analysis, correlation, and regression.

1 Introduction

A danger in statistics education is that we as educators focus on statistical knowledge (what to do, why to do it) and skills (how to do it) at the expense of instilling a real interest and desire to apply statistics [10]. Jim Albert, who has written multiple articles detailing baseball-related examples for use in statistics classes, argues that the popularity of baseball, and its suitability for statistical analysis, make it an excellent choice for demonstrating statistical methods in the classroom [1]. Students who follow baseball will naturally be more engaged in statistics if connections are drawn to their favorite hobbies, and even those who do not follow baseball specifically but are interested in sports generally can still readily apply statistical knowledge to their favorite sport. This is especially true in American higher education, as collegiate athletics are ubiquitous throughout American universities and are often core to an institution's identity, culture, and student life. By using datasets drawn from baseball or other sports, statistics becomes less of a chore that students must learn out of academic obligation and more of a new and exciting way to engage with their interests.

The wealth of available baseball data allows for deep statistical analysis in almost every aspect of the sport. Examples include use of neural networks [7] and multi-class classification [9] to predict pitch types in real time, ranking batters using Pareto optimization and weighted aggregation [12], estimating the conditional probability of winning a World Series game given the results of the previous game [4], evaluating draft prospects [5], and even analyzing the effect of team chemistry on wins [3]. This breadth of research means that data taken from baseball can be used to teach many different topics in statistics, such as basic data exploration [2], statistical inference [1], topics in linear regression [8], Poisson processes [6], etc.

This article presents a case study aimed at exploring relationships between pitching strategy (in particular, the characteristics and usage rates of the four primary pitch types: sliders, curveballs, changeups, and fastballs) and team performance. The data consists of team-level performance measures and characteristics and usage rates of different pitch types among teams in the Power Five Conferences. The 'Power Five' refers to five major athletic conferences (ACC, SEC, Big-10, Big-12, and PAC-12) in the NCAA's (National Collegiate Athletic Association) Division I that are known for their competitive athletics, with The NCAA being the organization that oversees these and other conferences, setting the rules and standards for collegiate athletics. Instructors can replicate our analyses in the classroom while teaching various statistics topics, including exploratory data analysis, correlation, and regression.

We considered two main response variables: (1) rating percentage index (RPI) and (2) whether a team made the playoffs; we intentionally chose one numeric and one categorical response variable so that we could use the data to illustrate a wide variety of statistics topics. RPI is a widely-used metric employed in collegiate sports to rank teams based on a multitude of factors, with strength of schedule being a crucial component. Ideally, a higher RPI rank signifies a “better” team. RPI can be obtained from public online resources (e.g. ncaa.com and d1baseball.com). In addition to having a high rank, making the playoffs is another natural indicator of a successful team. Out of 61 Power Five teams, only about half make the playoffs each year.

Explanatory variables consist of team-level pitch usage information, obtained through state-of-the-art data tracking technologies such as Trackman and Sportradar’s Synergy Sports technology software. Each pitch type is characterized in terms of a variety of averages, including exit velocity, strike (K) percentage, swing-miss rate, and usage rate. Exit velocity measures the average speed in miles per hour that the ball travels off the bat when hit by the batter. The strike percentage represents the frequency with which the pitch results in a strike, and the swing-miss rate denotes the percentage of times the opposing batter swings and misses at the pitch. Usage rate equals the proportion of times a particular pitch type is chosen. This data can be used in an effort to unravel the relationships between pitching strategy and overall team success. Using baseball topics as examples can also increase student interest and engagement in statistics and data science.

Our hope is that instructors are able to incorporate elements of this paper into their lesson plans for one or more classroom sessions in their introductory statistics courses. As a

reviewer pointed out, there is perhaps more here than is reasonable for a single sitting. With that in mind, we provide the following suggested roadmap for reading the paper. Most simply, an instructor could focus on sections 2 and 3 to introduce and explore the dataset, then sections 4.1, 4.3, and 4.6 as part of a basic coverage of linear regression. Depending on the desired depth of coverage, an instructor could include a discussion of multicollinearity (section 4.2), interactions (4.4), and model diagnostics (4.5). A similar roadmap for the logistic regression content would be sections 5.1, 5.3, and 5.6 for the basics and, if desired, section 5.2 for a discussion of multicollinearity, section 5.4 for interactions, and section 5.5 for model diagnostics. The dataset and all R code is included in an Appendix.

2 Dataset

The dataset, containing data from 61 teams, consists of 28 relevant variables that should be relatively easy for students to understand. See Table 1 for a list of variables with brief descriptions, and see the Appendix for a more detailed data dictionary.

Table 1

Abbreviated data dictionary containing description for variables in the dataset

teamAbbrevName	abbreviation for the college name
teamFullName	the full name of the college
BreakingBall Exit Velo	the velocity of the ball off the hitters bat when the team throws a breaking ball
BreakingBall Swing-Miss	the percentage of time that the batter swings and misses at the team's breaking ball
BreakingBall K%	the percentage of total breaking balls that result in a strike
BreakingBall Usage Rate	the percentage of time the team throws a breaking ball
SL Exit Velo	the velocity of the ball off the hitters bat when the team throws a slider
SL Swing-Miss	the percentage of time that the batter swings and misses at the teams slider
SL K%	the percentage of total sliders that results in a strike
SL Usage Rate	the percentage of time the team throws a slider
CB Exit Velo	the velocity of the ball off the hitters bat when the team throws a curveball

CB Swing-Miss	the percentage of time that the batter swings and misses when the team throws a curveball
CB K%	the percentage of total curveballs that results in a strike
CB Usage Rate	the percentage of time the team throws a curveball
CH Exit Velo	the velocity of the ball off the hitters bat when the team throws a changeup
CH Swing-Miss	the percentage of time that the batter swings and misses when the team throws a changeup
CH K%	the percentage of time the changeup results in a strike
CH Usage Rate	the percentage of time the team throws a changeup
FB Exit Velo	the velocity of the ball off the hitters bat when the team throws a fastball
FB Swing-Miss	the percentage of time that the batter swings and misses when the team throws a fastball
FB K%	the percentage of total fastballs that result in a strike
FB Usage Rate	the percentage of time the team throws a fastball
wins	number of wins for the team
losses	number of losses for the team
W/L ratio	ratio of wins to losses
RPI	rating percentage index; a quantity used to rank sports teams based upon team winning percentage, opponents winning percentage, and the opponent's winning percentage to capture the strength of team wins and losses while accounting for the difficulty of their schedule
Playoff Finish	where the team finished in the playoffs. Empty if they did not make it
win_percentage	the percentage of games the team won

The 'playoff finish' variable indicates the level of advancement achieved by each team in the college baseball postseason bracket. The college baseball playoff bracket consists of 64 teams each year, with 8 teams ultimately progressing to the college world series and competing for the championship. In the data frame, specific values are assigned to reflect

the team's performance: an N/A value signifies that the team did not qualify for the 64-team bracket, 49 denotes a loss in the 1st round, 33 indicates elimination in the second round, 17 signifies elimination in the third round, 9 represents the fourth round, 7 reflects elimination in the fifth round (first round of the college world series), 5 denotes elimination in the quarterfinals, 3 in the semifinals, 2 in the finals, and 1 signifies the team as the champion. Furthermore, the data frame includes additional variables such as 'wins' and 'losses,' which denote the number of games won and lost by the team, respectively. The 'W/L Ratio' is the ratio of wins to losses, the 'win percentage' represents the number of games won divided by the total games, and as previously mentioned, the 'RPI' serves as a ranking system, with a rank of 1 indicating the highest rank.

There are four variables for each pitch type (breaking ball, slider, curveball, changeup, and fastball). The variables are as follows: 'exit velocity,' which measures the average speed in miles per hour at which the ball travels off the bat when hit by the opposing team; 'strike percentage,' representing the frequency with which the pitch results in a strike; 'swing-miss rate,' indicating the percentage of times the opposing batter swings and misses at the pitch; and 'usage rate,' quantifying the proportion of times a particular pitch is chosen relative to all pitch types thrown by the team.

The pitching variables were obtained by the Texas A&M University baseball team. They employed both Trackman data and Sportradar's Synergy Sports technology to track and record pitch data for the 61 teams included in the dataset. Trackman uses radar technology that allows for precise measurement and tracking of pitch velocity, movement, and location for all 61 teams included in the dataset, while Sportradar's Synergy Sports technology provides comprehensive advanced statistical analysis and player performance tracking for every pitch in every game. The team success variables, such as `wins`, `losses`, and `RPI`, were sourced from a combination of the websites ncaa.com and D1baseball.com.

3 Exploratory Data Analysis

A natural first step in an analysis is to process the data by removing duplicates, handle missing values, standardize formats to the desired format, and address any other inconsistencies or errors within the dataset. Once the data have been processed, the next step that students should take is to perform exploratory data analysis (EDA). EDA involves a systematic approach to analyze and summarize the main characteristics of a dataset using summary statistics and visualizations. EDA aims to gain insights, identify patterns, detect outliers, and understand the underlying data structure.

In the initial stage of EDA, students should examine the summary statistics of each variable. These statistics show important measures such as the mean, median, mode, standard deviation, variance, range, and interquartile range (IQR). We provide simple R commands in the Appendix that can be used to reproduce all output and figures. For instance, Figure 1 presents the summary statistics of `RPI` in the year 2023, offering a concise overview of the variables' central tendencies. It also enables the identification of potential outliers, shown by the minimum and maximum values, while also highlighting the presence of any missing values (NA) for that particular variable if they exist. For RPI, it can be quickly noted that, on

average, teams in the data frame are ranked about 56th with the median team ranked 40th, which highlights that this variable is likely right-skewed. We also see that RPI ranges from 1 to 262 and the max of 262 is probably an outlier team that was ranked much worse than most other teams in our dataset.

Figure 1

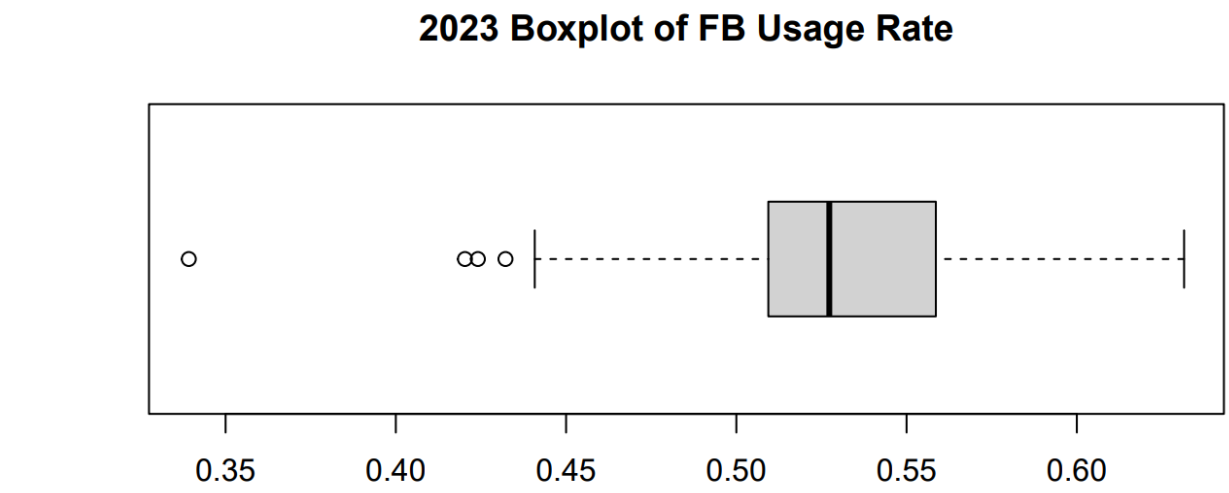
The output of `summary(RPI)` in R

```
Summary statistics for RPI :
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00   21.00   40.00   56.57   69.00  262.00
```

Figure 2 displays a basic box plot illustrating the usage of fastballs in 2023. The majority of the data points are concentrated between 0.44 and 0.64 , while four potential outliers are identified by points outside the box. Identifying outliers is essential because they may indicate unusual or extreme values that could significantly impact the analysis or interpretation of the data. Additionally, examining side-by-side boxplots for each year can reveal potential trends in the data. In Figure 3, a side-by-side boxplot of `BreakingBall Usage Rate` in 2021, 2022, and 2023 suggests an increasing trend in the usage of breaking balls over the years. This trend should prompt students to investigate further the reasons behind it and whether higher usage of breaking balls contributes to team success. Figure 2 also displays the histogram of fastball usage rate in 2023, showing a slightly left-skewed distribution with a single mode centered between 0.5 and 0.55 and an outlier between 0.3 and 0.35.

Figure 2

Boxplot and Histogram showing the spread of the `FB Usage Rate` variable in 2023


2023 Boxplot of FB Usage Rate
0.35
0.40
0.45
0.50
0.55
0.60

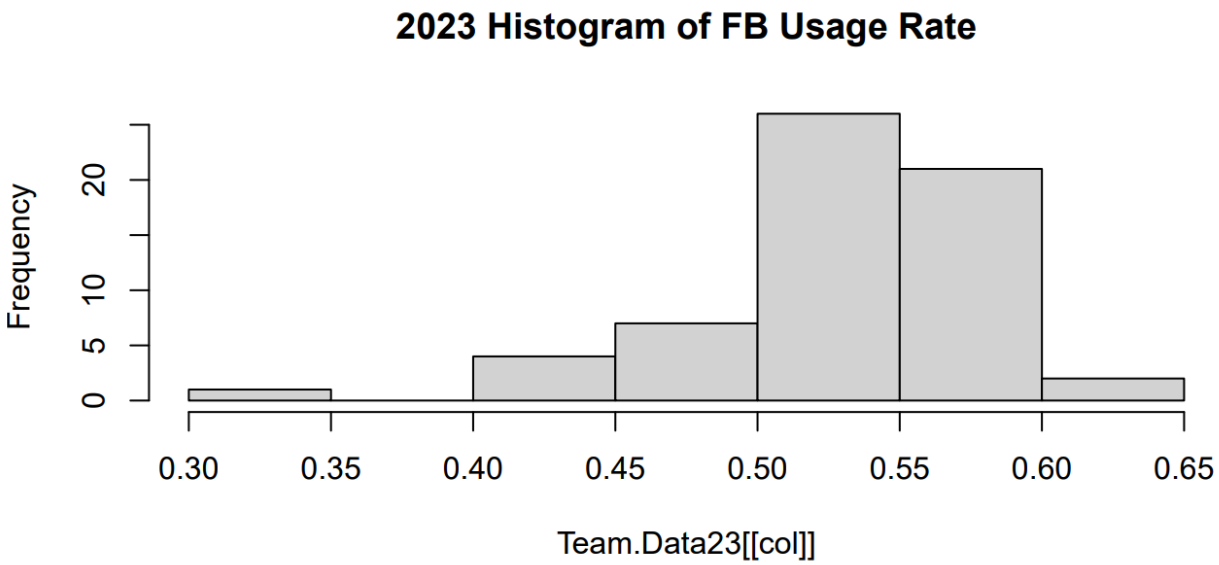
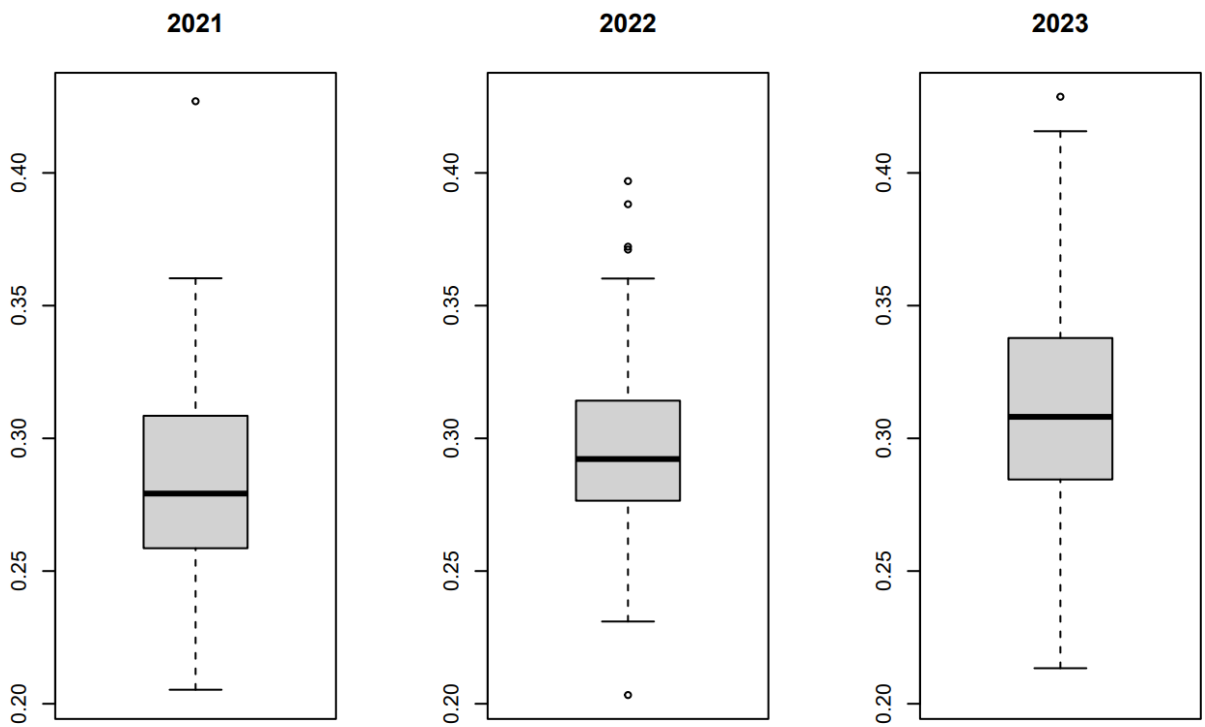

2023 Histogram of FB Usage Rate
Frequency
0
5
10
20
0.30
0.35
0.40
0.45
0.50
0.55
0.60
0.65
Team.Data23[[col]]

Figure 3

Side-by-side boxplot of the `BreakingBall Usage Rate` variable in 2021, 2022, and 2023



In addition to the above univariate summaries, there are many interesting bivariate relationships that can be explored with pairwise scatter plots and correlation matrices. In Figure 4, we see a scatter plot matrix of 4 of our variables. We can quickly notice there are no non-linear relationships present, but we have a high linear association between multiple variables such as `SL K%` and `SL Usage Rate`. Correlation matrices, as depicted in Table 2, provide valuable insights by displaying correlation coefficients. In Table 2, there is a strong correlation of 0.83 between `CB Swing-Miss` and `FB Swing Miss`, which makes sense considering that the dataset incorporates statistics for both sliders and curveballs under the variables for breaking balls. The correlation matrix raises the issue of multicollinearity between these variables, which can pose challenges in interpreting their individual effects during the modeling stage of our analysis. We will delve into this issue further in the following sections of this article.

Figure 4

Scatterplot Matrix showing relationships between `RPI` and 3 slider variables

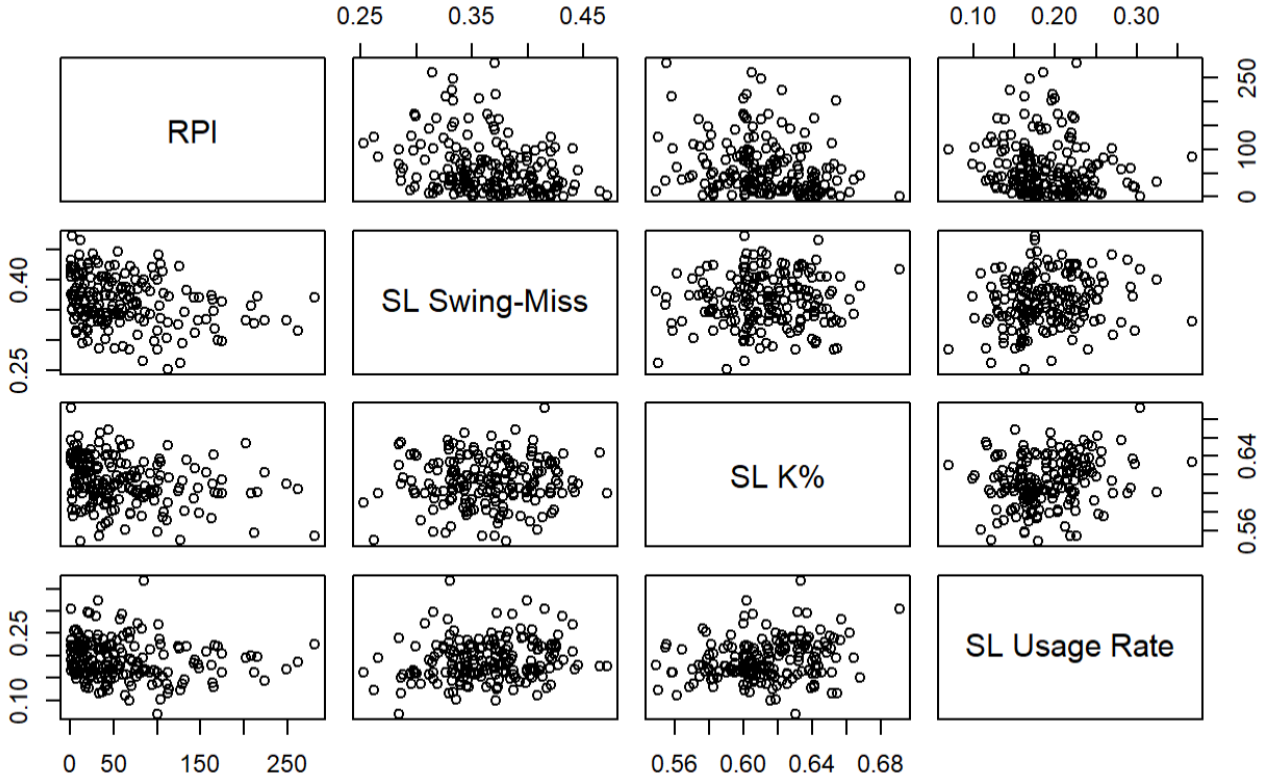


Table 2

Correlation Matrix of 9 variables in our data illustrating various levels of correlation and potential multicollinearity

	RPI	BreakingBall K%	BreakingBall Usage Rate	SL Usage Rate	CB Swing-Miss	CH Swing-Miss	FB Swing-Miss	FB K%
RPI	1.000	-0.36	-0.073	-0.130	-0.49	-0.360	-0.420	-0.29
BreakingBall K%	-0.360	1.00	0.380	0.400	0.24	0.030	0.170	0.24
BreakingBall Usage Rate	-0.073	0.38	1.000	0.700	0.27	-0.100	0.084	-0.15
SL Usage Rate	-0.130	0.40	0.700	1.000	0.28	-0.038	0.099	-0.21
CB Swing-Miss	-0.490	0.24	0.270	0.280	1.00	0.520	0.830	0.15
CH Swing-Miss	-0.360	0.03	-0.100	-0.038	0.52	1.000	0.270	0.24
FB Swing-Miss	-0.420	0.17	0.084	0.099	0.83	0.270	1.000	0.14
FB K%	-0.290	0.24	-0.150	-0.210	0.15	0.240	0.140	1.00

Also, from these scatter plots and correlation matrices, we can see which variables show association with RPI to get an idea of how we may be able to model team success. In Table 2, we can see that multiple variables have a slight negative correlation with RPI. For example, `FB Swing-Miss` has a -0.42 coefficient with RPI, indicating that a higher frequency of swing and misses at fastballs may be associated with a lower RPI. This relationship makes sense intuitively, and we should note `FB Swing-Miss` as a potential predictor variable in our model. This example also highlights the significance of comprehending the dataset and its variables because a lower RPI value (e.g., 1) signifies greater team success. Students should understand that, in this case, the proper interpretation of the -0.42 coefficient is that it indicates a slight positive correlation between having more fastball swing and misses and being a higher-ranked team.

4 Modeling RPI

4.1 Model Basics

We now illustrate the modeling of team success using linear regression on RPI. A natural first step toward building a model is to examine relationships between the response and all potential explanatory variables, one-at-a-time. Given that our response is numeric and all explanatory variables are numeric, we might begin by making a collection of scatterplots (like those in the ``pairs'' plot in Figure 4). While we do not show the resulting graphs here, we have included relevant code in the Appendix. When looking at the matrix of scatterplots, we want to focus on the graphs with `RPI` as one of the variables and try to find linear associations with other variables. As seen in Figure 4, `RPI` is negatively correlated with `SL Usage Rate' and 'SL Swing-Miss`. By inspection of all other pairwise scatterplots (see code in the Appendix), we see similar relationships involving `BreakingBall K%`, `BreakingBall Swing-Miss`, `CB Swing-Miss`, `CB K%, `CH Swing-Miss`, `FB K%`, and `FB Swing-Miss`. We therefore add these variables to our list of possible explanatory variables to use in a model.

Depending on the instructor's preferences, students could now be directed to examine whether the potential explanatory variables are approximately normally distributed. To illustrate, we will use the Shapiro-Wilks test, although less-formal visual assessments could be made with histograms or Q-Q plots. Each of `RPI', 'SL Usage Rate', `CB K%`, and `FB Swing-Miss` have Shapiro-Wilks p-values < 0.05, suggesting that they are not normally distributed. In our analysis, we applied Box-Cox transformations in an effort to make these variables more nearly-normally distributed; see the Appendix for details.

As our next step in the model-building process, we now examine a multiple regression model of 'RPI' against all the nine potential predictor variables that we identified based on the scatterplot matrix. Table 3 summarizes the results of this initial model. Notably, only `BreakingBall K%`, `CH Swing-Miss`, `FB K%`, and `FB Swing-Miss` are statistically significant, suggesting that not all nine variables are necessary to be included in the model.

Table 3

Summary of the initial linear regression model with `RPI` as the response variable.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	10.720	6.993	1.533	0.127
`BreakingBall Swing-Miss`	-4.881	5.818	-0.839	0.403
`BreakingBall K%`	-10.186	5.005	-2.035	0.043
`SL Usage Rate`	-2.130	2.043	-1.043	0.298
`CH Swing-Miss`	-5.574	1.858	-2.999	0.003
`CB Swing-Miss`	4.236	10.280	0.412	0.681
`FB K%`	-11.456	4.615	-2.482	0.014
`FB Swing-Miss`	7.322	3.462	2.115	0.036
`SL Swing-Miss`	0.495	4.967	0.100	0.921
`CB K%`	-4.097	2.961	-1.384	0.168

4.2 Multicollinearity and Variance Inflation Factors

This is an appropriate place to consider the issue raised earlier about correlation between the explanatory variables. Referring back to the correlation matrix in Figure 4 (and the larger scatterplot matrix involving all variables), we see that several of our explanatory variables are highly correlated with each other. One implication of correlated explanatory variables (also called "multicollinearity") is highly-variable / unstable regression coefficient estimates. In our case, for example, the standard error for the coefficient for 'CB Swing-Miss' is above 10. Another way to explore multicollinearity issues is with the Variance Inflation Factor (VIF) which quantifies the strength of association between one explanatory variable and all others in a model [11]. A common rule of thumb is that an explanatory variable is responsible for multicollinearity in a model if its VIF is greater than five. In our example, several variables have large VIFs (see Appendix). For the purposes of building an explanatory model, we do not want to use explanatory variables that provide

redundant information. Thus, in an effort to construct a more meaningful and interpretable model (as well as to remedy the numerical instability evident in our initial model estimates), we next filter our list of explanatory variables.

For each explanatory variable, we obtained other explanatory variables with which there is substantial correlation (based on a p-value < 0.05 for testing the null hypothesis of no correlation). This approach identified one cluster of four highly-correlated variables and multiple pairs of highly correlated variables. We used the strategy of selecting a single representative variable from each cluster of variables by regressing 'RPI' on each variable individually and choosing the one explanatory variable whose slope coefficient had the smallest p-value. However, we know that all of the breaking ball statistics encapsulate a combination of the slider and curveball statistics. Since we see that multiple slider, curveball, and breaking ball variables have multicollinearity, it is arguably most efficient to remove all slider and curveball variables from consideration. We will then regress `RPI` onto the remaining variables of each group and choose the single representative based on that result; see code in the Appendix for details. As an example, 'BreakingBall Swing-Miss' and 'FB Swing-Miss' are moderately correlated (r = .46). Of these, 'FB Swing-Miss' was most statistically significantly associated with RPI on its own, so we chose 'FB Swing-Miss' as the one representative variable for the effects of this cluster and removed the other variables from the data frame. After filtering explanatory variables in this way, all of the VIF values are well below 5, and thus we have resolved the multicollinearity issues.

A summary of the revised model is in Table 4. The adjusted R-squared value of approximately 0.39 indicates that about 39% of the variability in RPI can be explained by the four predictor variables. Given that we are only considering pitching-related explanatory variables, an R-squared of 0.39 seems reasonably high. Notice that the p-values for `Breaking Ball K%`, `CH Swing-Miss`, `FB Swing-Miss`, and `FB K%` are all below 0.05, indicating that each variable is statistically-significant. Because of this, we do not take any further model-selection steps.

Table 4

Summary of the revised linear regression model with `RPI` as the response variable after narrowing down the predictor variables

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	11.094	3.869	2.867	0.005
`BreakingBall K%`	-15.945	3.144	-5.072	0.000
`CH Swing-Miss`	-5.329	1.348	-3.954	0.000
`FB Swing-Miss`	7.124	1.430	4.980	0.000
`FB K%`	-10.980	4.278	-2.567	0.011

4.3 Model Interpretation

This would be a natural place for the instructor to discuss linear regression coefficient interpretation with the students. For example, the -10.980 coefficient for fastball strike percentage means that for a 1 percentage point increase in fastballs resulting in a strike, we would expect to see a decrease of 10.98 units in RPI. This suggests that college baseball

teams can improve team performance by throwing more fastballs for a strike (easier said than done, of course), and the interpretation of other coefficients follows a similar pattern.

4.4 Interactions

Depending on the instructor's interests, students could be directed to explore potential interactions. For example, Figure 7 shows the result of including an interaction between 'FB Swing-Miss' and 'BreakingBall K%'. The R output shown is the result of using an F-test to test for equivalence between two nested models (the model with the interaction compared to the nested model that does not have the interaction). The p-value is 0.038, indicating that the interaction between fastball swing-miss and breaking ball strike percentage is significant in the model. This suggests that the relationship between `FB Swing-Miss` and `RPI` is influenced by the value of `BreakingBall K%`. Specifically, the effect of fastball swing-miss on `RPI` depends on the level of breaking ball strike percentage. At lower values of `BreakingBall K%', the `FB Swing-Miss` variable has a greater effect on `RPI`, and as the value of `BreakingBall K%` increases, the effect of `FB Swing-Miss` on `RPI` diminishes. This pattern can be understood intuitively in the context of pitching strategy. When a team frequently achieves strikes with breaking balls, they establish a pattern that opposing batters begin to anticipate. Then, when the team throws a fastball, the batters, expecting a breaking ball, are more likely to be caught off guard, leading to increased swing-and-miss occurrences with fastballs. Note that adding the interaction raised the adjusted R-squared to 0.401, indicating an improvement in the model's explanatory power (but at the expense of more complex model interpretation).

Table 5

ANOVA table testing for interaction effects

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
No Interactions	178	168.4322	NA	NA	NA	NA
Interactions Included	177	164.4143	1	4.017941	4.325509	0.0389861

Table 6

Summary of the linear regression model with significant interactions included

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-89.988	48.753	-1.846	0.067
`FB K%`	-10.938	4.238	-2.581	0.011
`BreakingBall K%`	152.399	81.003	1.881	0.062
`CH Swing-Miss`	-5.150	1.338	-3.849	0.000
`FB Swing-Miss`	74.804	32.573	2.297	0.023
`BreakingBall K%`:`FB Swing-Miss`	-112.844	54.257	-2.080	0.039

4.5 Model Diagnostics

At this point, we turn our attention to model diagnostic checks. We consider the standard collection of diagnostic graphs in R: a residuals vs fitted values scatter plot and scatterplots of residuals vs each explanatory variable (for assessing the functional form of our model), a scatterplot of the absolute value of the square roots of residuals vs. fitted values (for assessing the assumption of constant error variance), leverage plots with Cook's Distance (for identifying outliers or influential values), and a Q-Q plot of the residuals (for checking whether the error terms are Normally distributed).

Figure 5 shows a scatterplot of residuals vs. fitted values. There is no apparent pattern to the plot, providing overall support for the model. Similarly, Figure 6 shows a scatterplot of residuals vs. fastball strike percentage, and the lack of any pattern provides evidence that this explanatory variable is modeled appropriately. The other scatterplots of residuals vs. explanatory variables look qualitatively the same (data not shown; see the Appendix). The Q-Q plot in Figure 7 shows that the sample quantiles align well with quantiles from a normal distribution, supporting the model's assumption of normality.

Figure 5

Residuals vs Fitted values scatter plot for our final linear regression model

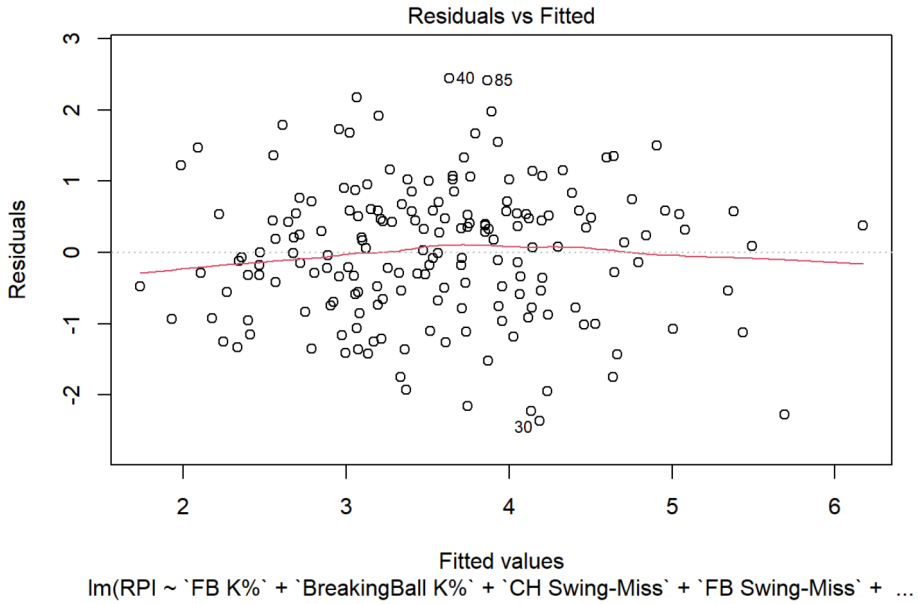


Figure 6

Scatter plot of the linear regression model residuals vs `FB K%`

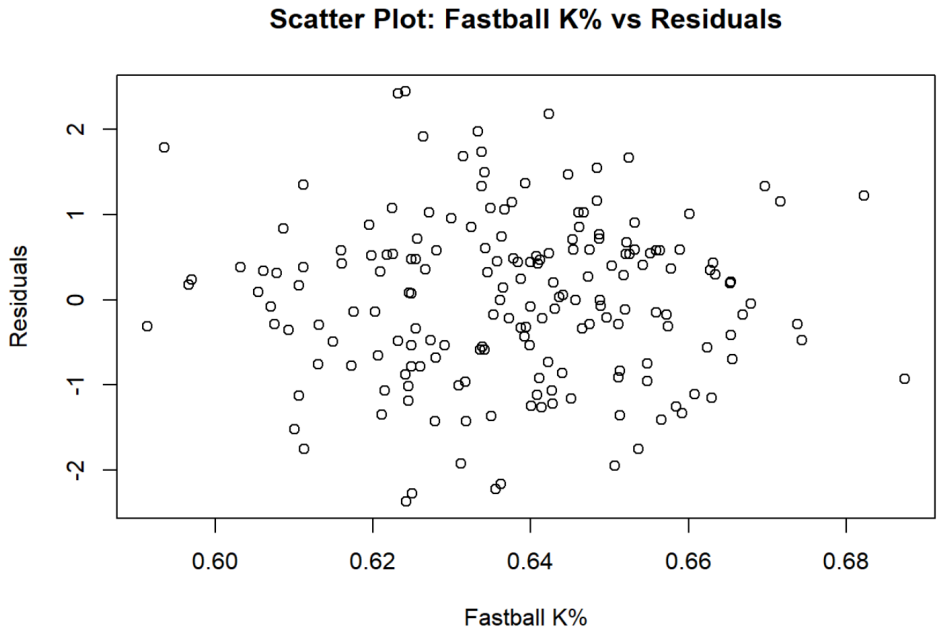


Figure 7

Q-Q plot of the standardized residuals from the linear regression model

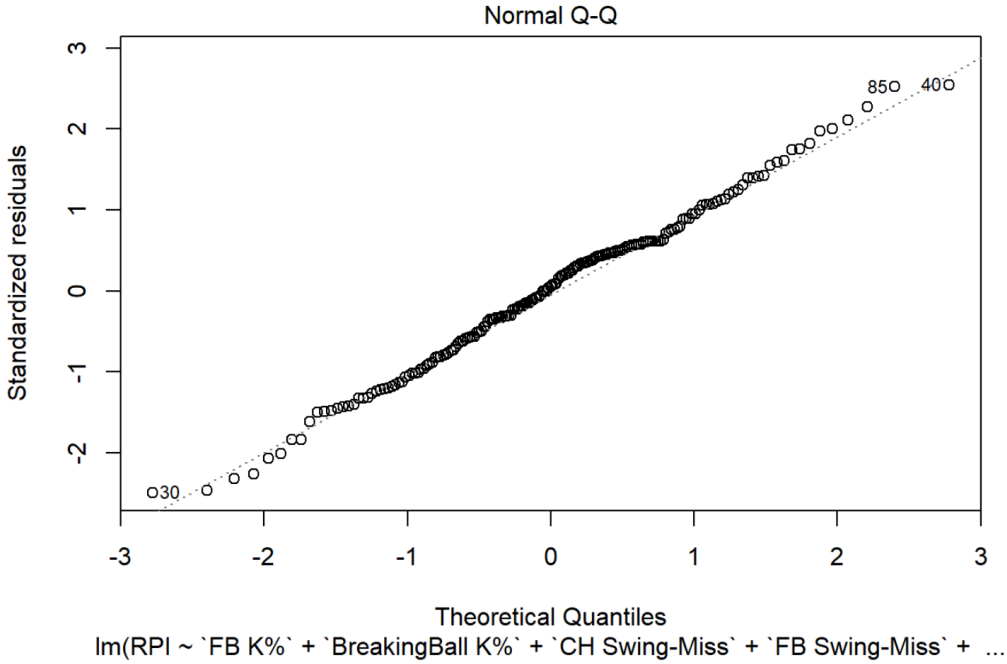
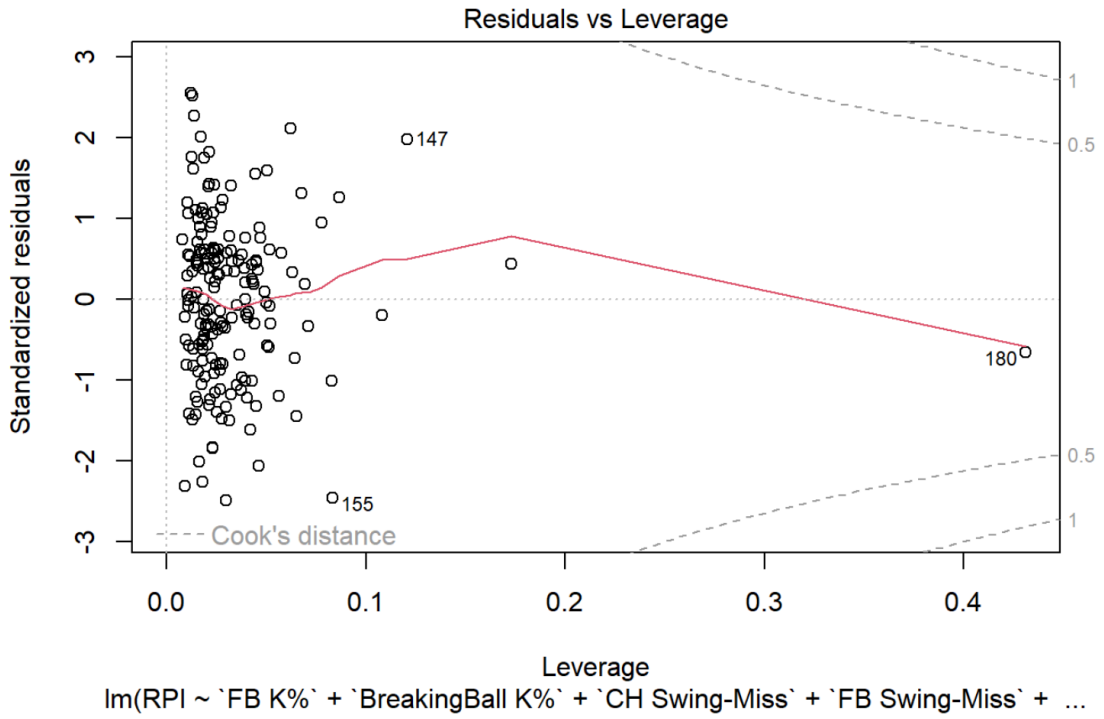


Figure 8 shows another of R's standard diagnostic plots for linear regression objects. It is a scatterplot of standardized residuals vs. leverage statistics, with contours overlaid that correspond to rules of thumb for identifying influential observations based on Cook's distance values. If any data points are beyond the dotted Cook's distance lines, it means these are influential observations that are substantially influencing the values of the model coefficients. In our case, there are no extreme Cook's distance values, so we are not concerned about influential points.

Figure 8

Residuals vs Leverage plot for the linear regression model



We do see a high-leverage observation (row 180) in Figure 8. Upon checking the data frame, we find that this row corresponds to the 2023 Wake Forest Baseball team. This is significant because the 2023 Wake Forest team was considered one of the best pitching teams in college baseball history, finishing the season with an RPI of 2. Based on inspection of their data, we observe that this team had significantly above-average fastball strike percentage, breaking ball strike percentage, changeup swing-miss rate, and fastball swing-miss rate, which are the four predictor variables in our model. The fact that the second-best rated team in 2023 and an all-time great pitching team demonstrated a substantial increase in the statistics that our model identified as significant gives us confidence that we have potentially created an accurate and real-world applicable model to predict college baseball team success.

4.6 Real-World Application

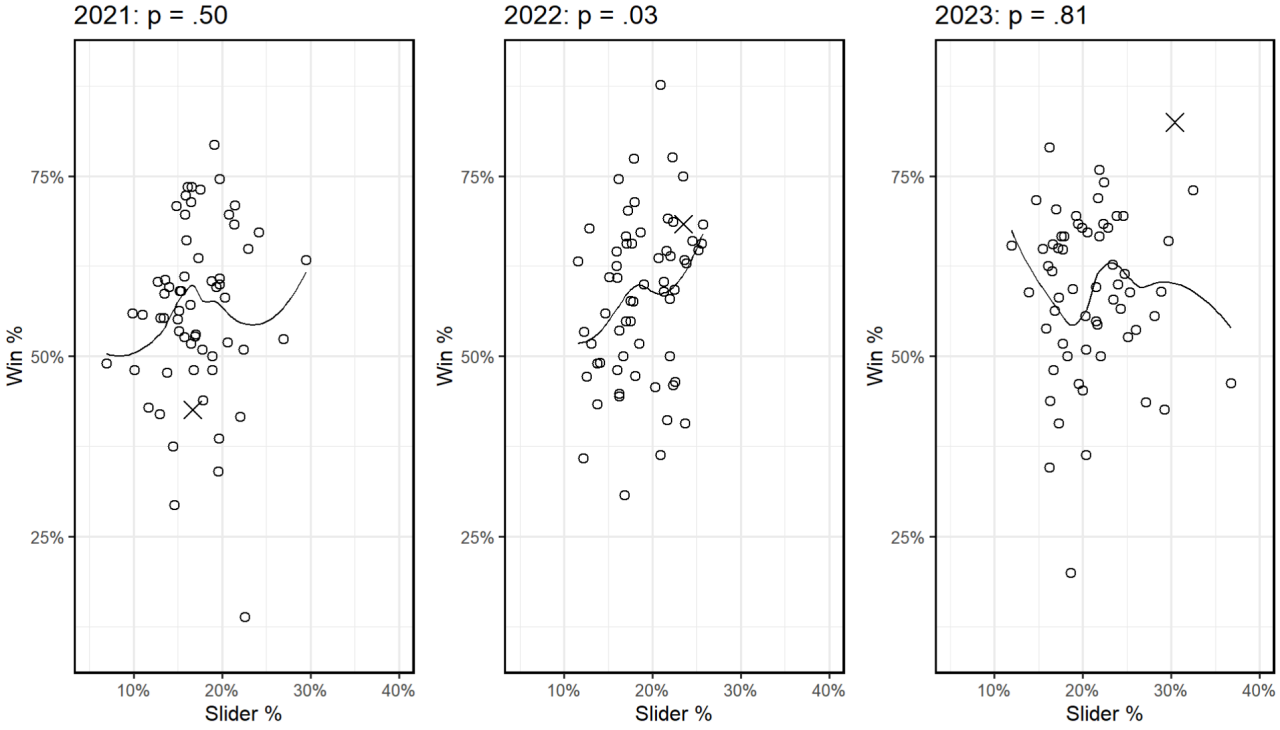
As a real-world application, students could suggest a strategy to college baseball coaches to focus on training their pitchers to throw more fastballs and breaking balls in the strike zone rather than trying to get batters to swing at pitches outside the strike zone. This observation is supported by the evidence from the 2023 Wake Forest Baseball team mentioned previously. Another notable takeaway is that we have empirically demonstrated that better overall pitching leads to a better team, confirming a notion that may be easily assumed but becomes both interactive and interesting when supported by underlying data.

Students should also be able to conclude from their model that the pitch usage variables for each team do not have a statistically significant effect on the team's performance over the three-year sample size. However, during exploratory data analysis, we observed that breaking ball usage, particularly slider usage, increased each year, while fastball usage decreased. Although it was initially hypothesized that this trend might be influential in our model, further analysis revealed that it is unnecessary and prompts a deeper dive into the data to explain this trend. Side-by-side scatter plots for slider usage in 2021, 2022, and 2023, as shown in Figure 9, are revealing. In the plots, the team plotted as an 'X' is Wake Forest, and the p-value in the title shows the statistical significance of the relationship each year. These plots seem to suggest that after the 2021 season, analytics teams may have discovered they could win more games by throwing more sliders.

Subsequently, the best teams in 2022 appear to have adopted this strategy and increased their slider usage, leading to a positive relationship with winning games. Then, in 2023, it appears this correlation diminished, and there was no longer a significant relationship between winning and throwing more sliders, potentially due to teams adjusting to this new philosophy and adapting their hitting strategy. The insights from Figure 9 explain why there is no significant correlation between pitch usage and higher rankings in our model, despite the observed correlation in 2021 and 2022. This evidence suggests that a key takeaway may be that when a strategic advantage is found in college baseball, teams only have a short period to exploit it before everyone else catches on and adapts their strategies accordingly.

Figure 9

Scatter plots of slider Percentage vs Win Percentage for 2021, 2022, and 2023 where the p-value in each title demonstrates the significance of the relationship (The 'X' is Wake Forest)



5 Modeling "Make Playoffs"

5.1 Model Basics

For any collegiate baseball team, the highest honor they can achieve is winning the College World Series. To do so, a team must play well enough in the regular season to make the 64 team tournament in the first place, then must win their super-regional tournament to be one of the 8 teams invited to the final rounds in Omaha, Nebraska, then beat 3 more teams before claiming the ultimate prize. Making the playoffs in the first place is already an impressive accomplishment, as you must be one of the best 64 teams out of the 299 Division I baseball programs.

Our team-level data includes the playoff finish of 63 Power-5 schools for 2021-23. The number indicates how far into the tournament the team progressed. 'Playoff Finish' = 1 is for the World Series champion (LSU for 2023), 2 is for the runner up (Florida), 3 is for teams eliminated in the semi-finals (TCU and Wake Forest), and so on; the higher the number, the earlier a team was eliminated, and rows with NA indicate teams that did not even make the playoffs.

Similar to our use of linear regression with the numeric response variable 'RPI', we now use logistic regression to model the probabilities for a binary response variable (whether a team does or does not make the playoffs) in terms of the same team-level explanatory

variables. Given the predictors $x_1, x_2, ..., x_n$, we will model the probability $p(x_1, x_2, ..., x_n)$ that a team makes the playoffs as $\log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n$ (i.e. the log of the odds that a team makes the playoffs is linearly related to our predictors).

To start, we first must convert the 'Playoff Finish' variable to a binary variable, 'in_Playoffs', and set it equal to 1 if a team has a value for 'Playoff Finish' (e.g., LSU, Florida) and 0 otherwise (e.g., UCLA). Next, we explore the relationship that each individual pitch usage variable has with making the playoffs. As an example, we start with 'Exit Velo BreakingBall.'

Table 7

Logistic Regression of 'in_Playoffs' with 'Exit Velo BreakingBall'

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	7.103	6.225	1.141	0.254
logit_data$`Exit Velo BreakingBall`	-0.084	0.074	-1.140	0.254

Based on the p-value for its regression coefficient, `Exit Velo BreakingBall` does not appear to be a significant predictor of `in_Playoffs`, so we will not use this explanatory variable in the final model. Students should repeat this for the other pitch usage variables and report which ones significantly contribute to the model for the log odds of `in_Playoffs`; these are 'BreakingBall Swing-Miss', 'BreakingBall K%', 'SL Swing-Miss', 'SL K%,' 'CB Usage Rate,' 'CB Swing-Miss,' 'CB K%,' 'CH Usage Rate,' 'CH Swing-Miss,' 'Exit Velo CH,' 'FB Swing-Miss,' and 'FB K%.'

We next fit the model with all of our potentially relevant predictors.

Table 8

Summary of Full Model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-39.647	11.385	-3.482	0.000
`BreakingBall Swing-Miss`	-2.205	17.191	-0.128	0.898
`BreakingBall K%`	24.284	39.925	0.608	0.543
`SL Swing-Miss`	5.290	12.597	0.420	0.674
`SL K%`	-2.112	26.600	-0.079	0.937
`CB Usage Rate`	-17.658	8.708	-2.028	0.043
`CB Swing-Miss`	8.188	29.533	0.277	0.782
`CB K%`	1.438	15.443	0.093	0.926
`CH Usage Rate`	-0.756	6.331	-0.119	0.905
`CH Swing-Miss`	5.265	5.078	1.037	0.300
`CH K%`	1.333	6.636	0.201	0.841
`Exit Velo CH`	-0.040	0.074	-0.539	0.590
`FB Swing-Miss`	13.803	18.586	0.743	0.458
`FB K%`	35.325	12.533	2.819	0.005

5.2 Multicollinearity and Variance Inflation Factors

As with our linear regression modeling of `RPI`, we see extreme standard errors as evidence of potential multicollinearity, and several explanatory variables have VIFs greater than 5 (see Appendix). By creating a correlation matrix with all of our predictors, we can identify those predictors that are highly correlated with each other. As before, we will start by dropping 'SL Swing-Miss,' 'SL K%,' 'CB Swing-Miss,' and 'CB K%,' as sliders and curveballs are both types of breaking balls, and since 'Breakingball Swing-Miss' and 'BreakingBall K%' are both statistically significant, they can represent the effect of all breaking balls. After refitting the model, none of the predictors have a VIF greater than 5.

Table 9

Summary of Model with Multicollinearity Fixed

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-40.233	11.266	-3.571	0.000
`BreakingBall Swing-Miss`	5.773	5.178	1.115	0.265
`BreakingBall K%`	25.074	8.434	2.973	0.003
`CB Usage Rate`	-16.395	5.680	-2.887	0.004
`CH Usage Rate`	0.238	5.681	0.042	0.967
`CH Swing-Miss`	6.063	3.525	1.720	0.085
`CH K%`	0.958	6.500	0.147	0.883
`Exit Velo CH`	-0.036	0.072	-0.494	0.622
`FB Swing-Miss`	18.236	7.630	2.390	0.017
`FB K%`	34.310	11.605	2.956	0.003

5.3 Model Revision

Table 9 summarizes the current model. Because there are several insignificant predictors in this model, we will remove some of them. There are of course many methods to decide which predictors to keep, such as forward selection, backwards selection, and stepwise selection, and several different criteria (p-values, AIC, BIC, etc.). For this example, we will use backwards selection and compare p-values; this results in a model with just `BreakingBall K%`, `CH Swing-Miss`, `FB Swing-Miss`, and `FB K%` as predictors. We note that there is a rich and instructive discussion to be had here about the general model-building process, and particularly the danger of "over-fitting." In an effort to limit the scope of this paper, we do not go further into that topic.

Table 10

Summary of Final Model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-41.626	8.590	-4.846	0.000
`BreakingBall K%`	25.134	8.167	3.077	0.002
`CB Usage Rate`	-17.103	5.642	-3.031	0.002
`CH Swing-Miss`	6.950	3.397	2.046	0.041
`FB Swing-Miss`	21.146	7.239	2.921	0.003
`FB K%`	34.618	11.262	3.074	0.002

5.4 Interactions

As we did with the model for `RPI`, we check for any significant interactions. As seen in Table 11, none of the first-order interactions are statistically significant, so we will not include any interactions in our model. Our final model for the log odds of `in_Playoffs` uses the predictors `BreakingBall K%`, `CH Swing-Miss`, `FB Swing-Miss`, and `FB K%`, as seen in Table 10.

Table 11

Summary of Model with Second-Order Interactions

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	346.361	216.076	1.603	0.109
`BreakingBall K%`	-463.742	343.630	-1.350	0.177
`CB Usage Rate`	13.856	273.252	0.051	0.960
`CH Swing-Miss`	-160.124	173.375	-0.924	0.356
`FB Swing-Miss`	19.853	389.803	0.051	0.959
`FB K%`	-646.381	352.105	-1.836	0.066
`BreakingBall K%:CB Usage Rate`	-154.797	303.410	-0.510	0.610
`BreakingBall K%:CH Swing-Miss`	233.481	173.174	1.348	0.178
`BreakingBall K%:FB Swing-Miss`	-601.405	401.133	-1.499	0.134
`BreakingBall K%:FB K%`	863.191	546.724	1.579	0.114
`CB Usage Rate:CH Swing-Miss`	42.590	123.693	0.344	0.731
`CB Usage Rate:FB Swing-Miss`	64.259	255.608	0.251	0.802
`CB Usage Rate:FB K%`	50.088	398.264	0.126	0.900
`CH Swing-Miss:FB Swing-Miss`	-177.810	156.554	-1.136	0.256
`CH Swing-Miss:FB K%`	96.390	214.485	0.449	0.653
`FB Swing-Miss:FB K%`	646.854	533.573	1.212	0.225

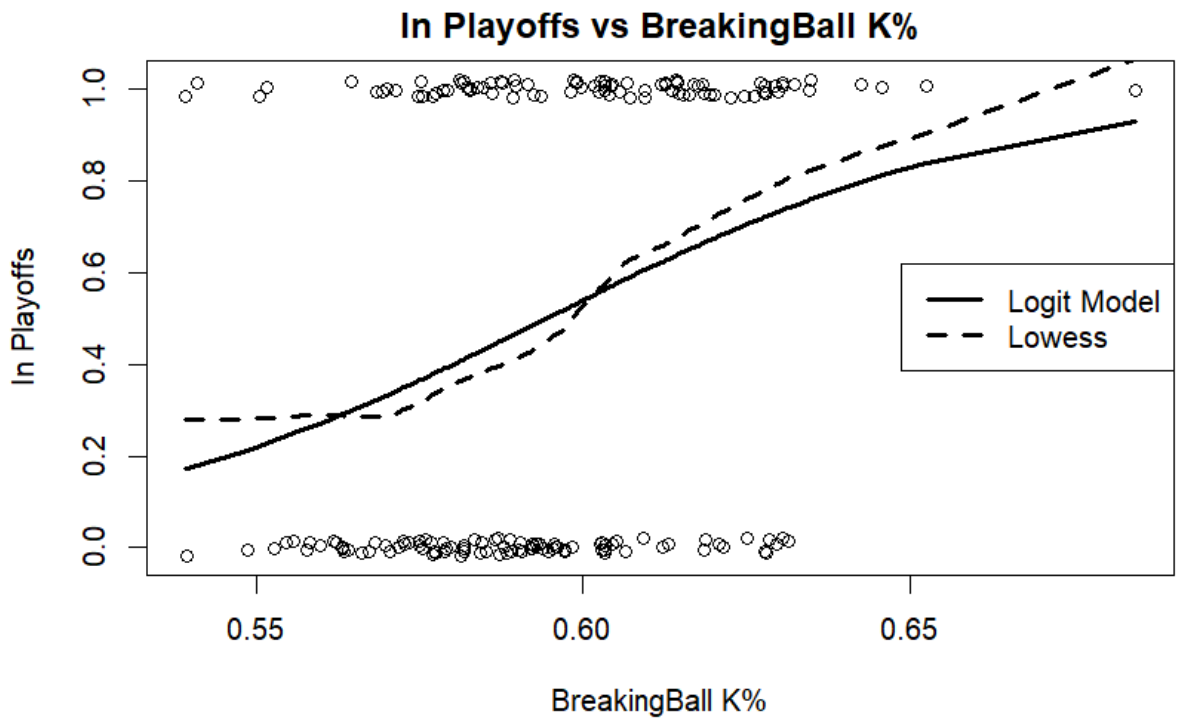
5.5 Model Diagnostics

For the purposes of diagnostics with our logistic regression model, we plot fitted probabilities of 'Make Playoffs' against our explanatory variables (referred to as "marginal model plots") and compare them with nonparametric (lowess) fits [11]. While there are residuals and associated plots that can be made with logistic regression models, they are notoriously difficult to interpret, and we argue that marginal model plots of estimated probabilities are more straightforward to teach and use in the classroom.

To illustrate the use of a marginal model plot, we fit a logistic regression model using 'BreakingBall K%' as the sole explanatory variable. Figure 10 then shows our diagnostic plot for the 'BreakingBall K%' variable. The y-axis values (zeros and ones) have been jittered randomly to aid visualization. The solid curve corresponds to the fitted probabilities of 'Make Playoffs' from our model, while the dashed curve is a nonparametric (lowess) curve fit to the raw observations. If our model is "valid," these two curves should largely agree. Judging whether the curves "agree" is unavoidably subjective. In this example, while the model-based curve is somewhat below the nonparametric curve for extreme values of `BreakingBall K%`, there is overall agreement, so we would conclude that the model is incorporating this explanatory variable appropriately.

Figure 10

Diagnostic Plot for Logistic Model with 'BreakingBall K%' only

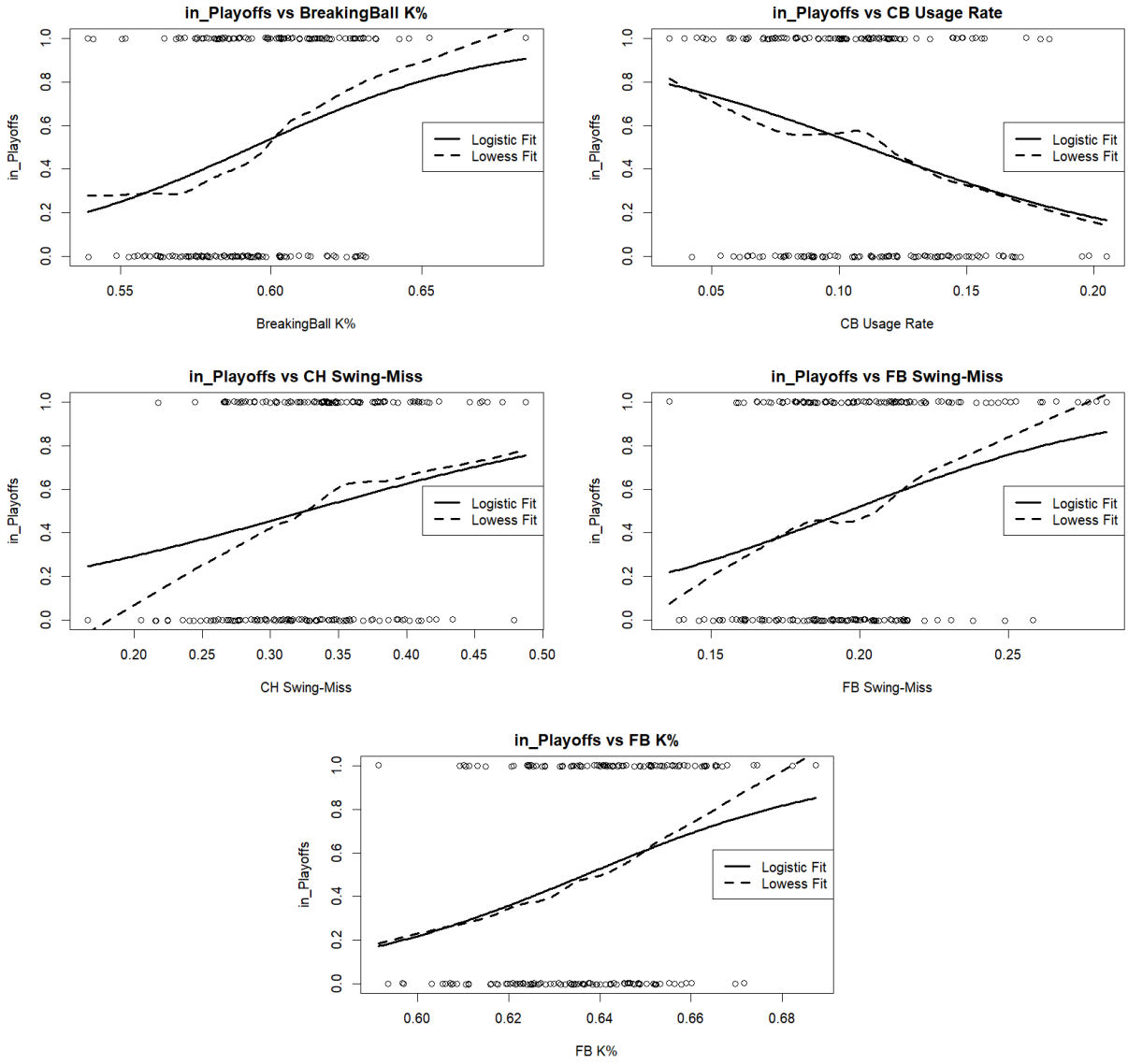


To create marginal model plots for a model with multiple predictors, we will make one of these plots for each explanatory variable. For example, for 'BreakingBall K%', we will obtain

fitted probabilities for 'Make Playoffs' when all other explanatory variables are set to their average values. We repeat this process for every other predictor in the final model (Figure 11). If the predicted curves are all close to the nonparametric curves, then we conclude that our model is "valid."

Figure 11

Diagnostic Plots for each Predictor in our Final Model



There is some substantial deviation between curves in the tails, particularly for the variables `CH Swing-Miss`, `FB Swing-Miss`, and `FB K%`. This implies that our model may be under- or over-estimating the effect certain variables have on the odds of making the playoffs. Thus, the model is clearly imperfect. To remedy this, we could try to transform our predictors, or we might search for additional explanatory variables to include in the model. We leave this as an exercise for the reader (and his / her students). For the purposes of

what follows, we refer to our logistic regression model as an *approximate* representation of 'Make Playoffs'.

5.6 Model Interpretation

As with the linear regression model for `RPI`, it is instructive to consider the interpretation of the logistic regression model. For example, the coefficient for `FB K%` is 34.618, meaning that a one-percent increase in fastballs resulting in strikes is associated with a $e^{.34618} = 1.414$ odds ratio of making the playoffs.

We found that the most significant factors that increase a team's odds of making the playoffs are 'BreakingBall K%,' 'CB Usage Rate,' 'CH Swing-Miss,' 'FB Swing-Miss,' and 'FB K%.' The exit velocity for all pitch types was not statistically significant. Usage rate similarly did not contribute substantially (only curveball usage rate made it into our final model). This suggests that winning games is more about pitching well on average than about how frequently you use different pitch types. Interestingly, 'CB Usage Rate' did have a significant effect on the odds of making the playoffs, but it had a negative coefficient. This suggests that teams that relied on curveballs were less likely to make the playoffs. We also note that the coefficient of 'CH Swing-Miss' is lower in magnitude than the other predictors, and its p-value is much higher, implying that changeups have a smaller effect on making the playoffs than breaking balls and fastballs. As shown previously, there is some disagreement between our model and a non-parametric fit, implying that the model may be undervaluing the effect of low values of `CH Swing-Miss` on the odds of making the playoffs. It is possible for students to draw opposite conclusions from the same data and analysis. This ambiguity is part of what makes this data set a good tool for teaching: it encourages students to critically think about their data in a context that they are interested in, instead of following a checklist of statistical tests on data from a context they simply are not engaged with.

6 Discussion

The baseball dataset provided contains two main response variables, `RPI` and `Make Playoffs`, as well as 20 potential explanatory variables. The dataset can be used by instructors to illustrate a number of topics in exploratory data analysis and regression. These include analyzing summary statistics, exploring box plots and histograms, viewing correlation matrices, modeling with linear regression, modeling with logistic regression, finding interaction effects, dealing with multicollinearity, addressing violations of the normality assumption, and how to build models.

One limitation is the fact that only 61 college baseball teams use the technology necessary for these statistics to be tracked. We would prefer, ideally, to have data on all college baseball teams. Another limiting factor is that this technology was introduced only in 2021, resulting in a dataset with only three years of data. Still, the data provide real-world insights in a context that many students will find interesting. This allows instructors to better engage and interest students when demonstrating statistics concepts. Additionally, the dataset facilitates an easily interpretable process for performing exploratory data analysis and regression.

For future work, it would be beneficial to gather further data on each specific type of pitch, such as baseball spin rate and break angle. This additional information could help explain how teams can improve their ability to throw more strikes and increase opposing batter swing-miss.

References

1. J. Albert, A Baseball Statistics Course, *Journal of Statistics Education* vol. 10 no. 2 (2002).

2. J. Albert, Baseball Data at Season, Play-by-Play, and Pitch-by-Pitch Levels, *Journal of Statistics Education* vol. 18 no. 3 (2010).

3. S. Brave, R. A. Butters, K. Roberts, Uncovering the Sources of Team Synergy: Player Complementarities in the Production of Wins, *Journal of Sports Analytics* 5 (2019) 247–279.

4. D. Chance, Conditional Probability and the Length of a Championship Series in Baseball, Basketball, and Hockey, *Journal of Sports Analytics* 6 (2020) 111–127.

5. C. Conforti, Christian M., R. Crotin, J. Oseguera, Major League Draft WARs: An Analysis of Wins Above Replacement in Player Selection, *Journal of Sports Analytics* 8 (2022) 77–84.

6. M. Huber, A. Glen, Modeling Rare Baseball Events — Are They Memoryless?, *Journal of Statistics Education* vol. 15 no. 1 (2007).

7. J. S. Lee, Prediction of Pitch Type and Location in Baseball Using Ensemble Model of Deep Neural Networks, *Journal of Sports Analytics* 8 (2022), 115–126.

8. F. Samaniego, M. Watnik, The Separation Principle in Linear Regression, J*ournal of Statistics Education* vol. 5 no. 3 (1997).

9. G. Sidle, H. Tran, Using Multi-class Classification Methods to Predict Baseball Pitch Types, *Journal of Sports Analytics* 4 (2018) 85–93.

10. R. Snee, What's Missing in Statistical Education?, *The American Statistician*, vol. 47, no. 2, (1993) 149–54.

11. S. Sheather, *A Modern Approach to Regression with R*, Springer, 2005.

12. S. Wulf, W. P. De Silva, A Multi-criteria Approach for Evaluating Major League Baseball Batting Performance, *Journal of Sports Analytics* 8 (2022) 85–98.