Raymond Provost

Ms. Kiernan

Linear Algebra

February 14, 2023

Linear Algebra in “Moneyball”

           Moneyball is a baseball strategy that is rooted in sabermetrics, an idea developed by Bill James in the 1980s that statistics could provide a predictive value in analyzing the performance of baseball players. For a century, baseball had been viewed as a game of mystery, where chance had much to do with one’s success. Bill James sought to break it down to a science using statistics to learn objective truths concerning baseball outcomes. Nearly two decades later, James’ framework was adapted by Paul DePodesta, Billy Beane, and the Oakland Athletics organization to compete with the big-budget teams in Boston, Los Angeles, and New York with the budget of a small-market club. After implementing what was known as the Moneyball strategy, Oakland had their most successful season in years, winning the division and making the playoffs in the process.

           Rather than assembling the cheapest possible winning team, I sought to evaluate the current state of MLB contracts using concepts from Linear Algebra and Machine Learning. One of the core concepts of Machine Learning is Linear Regression. The idea behind Linear Regression is that one independent variable can be used to predict the value of a dependent variable. In a simple Linear Regression, the dependent variable axis is plotted as a function of the independent variable. After all, points have been plotted, a line of best fit (mx + b) is drawn using the Least-Squares method. The Least-Squares method creates a line of best fit such that the total squared distance from each point to the line is minimized. In my Linear Regression model, the independent input variables were players' stats from the 2022 season, and the dependent variable was their salary. The stats used were carefully selected as stats that measured a player’s ability within an isolated context, or stats that do not depend on the other players on the field. The primary difference between this model and a simple Linear Regression model is that there is now more than one input variable. This means that while the data can still be fit with a linear function of the first degree, the fit must be of the same dimension as the input variables (m1x1+ m2x2+ m3x3+…+mnxn + b). In other words, with five input variables, there would be a five-dimensional object of best fit. This type of Regression model is known as Multiple Linear Regression. The challenge of using Multiple Linear Regression as opposed to Simple Linear Regression is that the prediction of the output variable can no longer be calculated by finding the new Y-coordinate that corresponds to the original X-coordinate. However, the principle of minimizing the distance between the point and the object of best fit is still applicable. To do this, it is necessary to think of the object of best fit as a vector space, a subspace of Rn, n being the number of input variables. As well as the points as Vectors in Rn. To calculate the Y-coordinate of the vector in the vector space of best fit closest to the original point, an orthogonal projection of the point onto the vector space must be performed. After obtaining the new projected vector, the Y-coordinate will either be higher or lower than the original, depending on if the point was above or below the fit. The closer a player’s predicted salary is to his actual salary, the more accurate his contract is.

           Generally, this model works well for evaluating the contracts of MLB players, however, it has some flaws that appear in extreme cases. As expected, the most underpaid players are players still on league minimum contracts that had breakout seasons in 2022. However, since the league minimum is $700,000, there is little to no data below this point on the graph. Because of this, the fit is over-adjusted, and the y-intercept can become negative. In some cases, this causes players whose contracts are close to the league minimum to have salaries projected in the negatives. In addition, this Linear Regression model was performed on only one dataset. Typically in machine learning, a training dataset is used, followed by a testing dataset where predictions are made. This ensures that the model can be generalized to all input data of the same form, rather than becoming overfit to one dataset.

Bibliography

“DataFrame#.” DataFrame - Pandas 1.5.3 Documentation, https://pandas.pydata.org/docs/reference/frame.html.

“Major League Leaderboards " 2022 " Batters " Dashboard: Fangraphs Baseball.” Major League Leaderboards " 2022 " Batters " Dashboard | FanGraphs Baseball, <https://www.fangraphs.com/leaders.aspx?pos=all&stats=bat&lg=all&qual=y&type=8&season=2022&month=0&season1=2022&ind=0&team=0&rost=0&age=0&filter=&players=0>.

“Major League Leaderboards " 2022 " Pitchers " Dashboard: Fangraphs Baseball.” Major League Leaderboards " 2022 " Pitchers " Dashboard | FanGraphs Baseball, <https://www.fangraphs.com/leaders.aspx?pos=all&stats=pit&lg=all&qual=100&type=8&season=2022&month=0&season1=2022&ind=0&team=0&rost=0&age=0&filter=&players=0&startdate=&enddate=>.

Mikulik, Vladimir. “Why Linear Regression Is a Projection.” Medium, Medium, 19 Dec. 2017, <https://medium.com/@vladimirmikulik/why-linear-regression-is-a-projection-407d89fd9e3a>.

Paruchuri, Vik. “Using Linear Regression for Predictive Modeling in R.” Dataquest, 8 Feb. 2023, <https://www.dataquest.io/blog/statistical-learning-for-predictive-modeling-r/>.

“The Linear Algebra behind Linear Regression.” GoDataDriven, 29 Oct. 2020, https://godatadriven.com/blog/the-linear-algebra-behind-linear-regression/.