Baseball Statistics

[Enter your name here]

[Enter name of the institution here]

Dependent variable is the salary paid to a player of a baseball team. Baseball is the national game of the US and a large number of stakeholders are interested in this game. Various stakeholders are interested in this variable especially the owners of team and investors. The dependent variable is explained by scores, winning percentage, batting average, home runs, runs, earned run average and pitching saves.

The primary independent variable is pitching saves defined by the number of saves, or percentage of saving opportunities converted successfully. This is the most important variable because this one act of saving the runs is the most critical to the results of any match. Hitting home run is a chance item that may or may not happen in every match. Payroll is also a variable that is not directly associated with the performance in the game or match. Thus, we have considered pitching saves as the most critical independent variable among all. The positive aspect in measuring winning percentage is that this variable is measured without any ambiguity (Porter & W.Scully, 1982). There are more wins reported by the teams whose players have saved more runs. (Hall, Szymanski, & S.Zimbalist, 2002).

Salaries = Saves + winning percentage + runs

Where,

Payroll is the amount of salary paid to the player in a month. This is chosen as the dependent variable because salary is very important to the player as well as to the team. Teams pay salaries to players so they can help them win matches. More effective players receive higher salaries and this effectiveness is shown by the number of runs scored or number of saves made.

Winning percentage is defined as the number of matches won by a team as a percentage of total matches played over the period of study.

Saves show the saves per match made by the pitcher over the period of study.

The usefulness of this study is that it will help various baseball players to identify the factors affecting their salary structure. The dependent variable is the amount of salary that an individual player will receive every month. This amount is adjusted as a percentage of average salaries paid by a team in the specified period (Hassan, n.d.).The study is important because salaries are often the biggest expenses made by a team. The teams will also determine the increase in salaries of certain players on the bases of their performance. There are three independent variables namely score, winning percentage and saves. Score is defined as the runs scored by a player over a period of one-month, winning percentage is calculating by taking a percentage of matches won as compared to the total matches played and saves are recorded as the shots saved by a player on position other than the pitcher. The expected sign of the independent variable is positive as the high scoring players will be offered a higher salary. Similarly, the teams who have higher winning percentages are expected to pay their players a higher salary (T.Mangine, et al., 2013).

The information regarding salaries paid to players is gathered from the websites of various teams. These websites also hold the record of team’s winning percentage and saves made by renown players (MLB.com, 2019). The data for salaries will be presented in dollars whereas the data of saves and runs will be presentenced in numbers. The data for winning will be presented in form of percentages. Since all the variables are having different units, the logs of all variables are taken to make the analysis more logical and comprehendible. The limitation to this data is that the websites update the data regularly every 5 years and the previous data is deleted from the databases. We have gathered only three years data i.e. for 2017,2018 and 2019.

**The Model**

Salary = a + b1\*score + b2\* winning percentage+ b3 \* saves

When regression is run, the r square is calculated as 0.84 which means that the independent variables in the model explain 84% variation in the dependent variable. The remaining 16% variation is defined by the error term and is not defined by the variables included in this model. The strongest independent variable is the winning percentage as depicted by the highest value of its coefficient of 5.67 which shows that every unit change in the salary of baseball players increases the winning percentage by 5.67 units. The coefficient of scores comes to 3.02 and that of saves is 1.02. This means that the least impact on salaries is shown by the variable saves. The regression analysis is prone to certain assumptions, variables considered in the regression analysis are assumed to have a linear relationship with each other. The linear relationship can be seen by making a scatter diagram involving a dependent and an independent variable. All variables that are included in regression analysis are assumed to be multivariate normal which can be checked with the help of a histogram or Q-Q plot. A goodness of fit test can also be used to check the normality of these variables. A model in which data is not normally distributed, some transformation should be applied to data e.g. log transformation. One of the most important assumptions is the existence of multicollinearity in the data. This means that the independent variables used in the data have strong relationships within themselves and such relationships do not allow the researcher to study the relationship between dependent and independent variables properly. To test for the presence of this problem, any of the three methods can be used. A correlation matrix is developed to observe the correlation scores between all the independent variables in the form of coefficients. For a satisfactory result, scores on the correlation matrix should be very close to zero. The tolerance method measures if there is a significant influence of one independent variable on the other variables. If the tolerance score for any variable is less than 0.01, there is certainly some existence of multicollinearity. The variance inflation factor is the reciprocal of the measure of tolerance. A score of more than 100 in this measure shows that multicollinearity exists. To solve the problem of multicollinearity, some of the independent variables can be removed from the analysis or some new variables can be added to the analysis. In the above analysis there is a problem of multicollinearity that exists between winning percentage and the other two variables namely saves and scores. It is suggested that some other variables may also be added to the analysis to cure this problem.

The adjusted R square is the value that shows power of a regression analysis. The value of 0.76 in this model shows that 76 % variation in the dependent variable is shown by the independent variables. The rest 24% variation is account ted for by the error term or other variables that have not been included in the model. The choice of independent variables is reasonable as a majority of variation is defined by the variables included in the model. If this value is lower than 50%, there is a need to include or exclude variables from the model.

The F test of overall significance shows whether the linear regression model provides a suitable fit to the data than a model that contains no independent variables. F tests can evaluate multiple values at the same time which allows that fits of more than one models can be judged at the same time. This feature is not available with the t- tests which can analyze a single value at a time. The p-value is 0.002 which shows a good fit for the model as this test is conducted at a significance level of 0.05 and p- value is less than the level of significance.

The t- tests for the scores variable show that the probability value is less than 0.05 which shows that there is no significant difference between the mean score of the variable scores and the score of same variables from the population. The null hypothesis is accepted regarding no significant difference between the sample and population values.

The value of t-tests for the variable winning percentage shows that p-value is 0.08 so in order to accept that this value is not significantly different from the value of same variable in the population, we have used a level of significance at 0.10. The null hypothesis is accepted regarding no significant difference between the sample and population values for the said variable.

The value of t tests for the variable saves shows that the probability value is 0.004 which is significant at 0.05 level of significance and shows that the value of this variable is not significantly different from the value of similar variable from the population. This shows that for this variable, there is no significant difference between the sample score and the population scores.

Yes, the signs of all the independent variables are found to be positive as depicted because the more scores will mean more winning percentage and ultimately more salaries to the players.

# References

Hall, S., Szymanski, S., & S.Zimbalist, A. (2002). Testing causality between team performance and Payroll: The cases of major league Baseball and Soccer. *Journal of Sports Economics*.

Hassan, S. (n.d.). *https://pdfs.semanticscholar.org/6f39/892055920f0a7855fbc82f50db2f5b46bcf8.pdf*. Retrieved from https://pdfs.semanticscholar.org: https://pdfs.semanticscholar.org/6f39/892055920f0a7855fbc82f50db2f5b46bcf8.pdf

MLB.com. (2019, July 31). *https://www.mlb.com/orioles/official-information*. Retrieved from https://www.mlb.com: https://www.mlb.com/orioles/official-information

Porter, P. K., & W.Scully, G. (1982). Measuring managerial efficiency: The Case of Baseball. *Southern Economic Journal*, 642-650.

T.Mangine, G., J.R.Hoffman, Vazquez, J., Pichardo, M., Fragala, M. S., & Stout, J. R. (2013). Predictors of Fielding Performance in Professional Baseball Players. *International Journal of Sports Physiology and Perfromance*, 510-516.