DATA621 | Project I

Abdellah AitElmouden | Gabriel Abreu | Jered Ataky | Patrick Maloney

2/12/2021

Abstract

To see how regression will help us evaluate baseball team performance, we will explore, analyze and model a historical baseball data set containing approximately 2200 records, to determine a team's performance based on statistics of their performance. Each record represents a professional baseball team from the years 1871 to 2006 inclusive, and the data include the performance of the team for the given year, with all of the statistics adjusted to match the performance of a 162 game season.

While correlation does not equal causation it is suggested that a focus on some of the variables such as a focus on either single hits or triple or more hits to the exclusion of doubles might be worth pursuing. Also the data suggests that a focus on home runs allowed may not be worth giving up a number of more normal hits.

.....To add more here.....

Introduction

Because baseball is so numbers-heavy, there are many different statistics to consider when searching for the best predictors of team success. There are offensive statistics (offense meaning when a team is batting) and defensive statistics (defense meaning when a team is in the field). These categories can be broken up into many more subcategories. However, for the purpose of the this project we will use the available data to build a multiple linear regression model on the training data to predict the number of wins for the team.

Data Used

the data was provided in csv file. The files contain approximately 2200 records. Each record represents a professional baseball team from the years 1871 to 2006 inclusive. Each record has the performance of the team for the given year, with all of the statistics adjusted to match the performance of a 162 game season.

VARIABLE NAME	DEFINITION	THEORETICAL EFFECT
INDEX	Identification Variable (do not use)	None
TARGET_WINS	Number of wins	Outcome Variable
TEAM_BATTING_H	Base Hits by batters (1B,2B,3B,HR)	Positive Impact on Wins
TEAM_BATTING_2B	Doubles by batters (2B)	Positive Impact on Wins
TEAM_BATTING_3B	Triples by batters (3B)	Positive Impact on Wins
TEAM_BATTING_HR	Homeruns by batters (4B)	Positive Impact on Wins
TEAM_BATTING_BB	Walks by batters	Positive Impact on Wins
TEAM_BATTING_HBP	Batters hit by pitch (get a free base)	Positive Impact on Wins
TEAM_BATTING_SO	Strikeouts by batters	Negative Impact on Wins
TEAM_BASERUN_SB	Stolen bases	Positive Impact on Wins

VARIABLE NAME	DEFINITION	THEORETICAL EFFECT
TEAM_BASERUN_CS	Caught stealing	Negative Impact on Wins
TEAM_FIELDING_E	Errors	Negative Impact on Wins
TEAM_FIELDING_DP	Double Plays	Positive Impact on Wins
TEAM_PITCHING_BB	Walks allowed	Negative Impact on Wins
TEAM_PITCHING_H	Hits allowed	Negative Impact on Wins
TEAM_PITCHING_HR	Homeruns allowed	Negative Impact on Wins
TEAM_PITCHING_SO	Strikeouts by pitchers	Positive Impact on Wins

Data exploration

Variable Selection

Outliers

Correlations among predictors

Data Preparation

Build Models

Select Model

Appendix

References