

DATA 621: Homework 1 (Group 2)

Moneyball Linear Regression

Group Members: *Alicia Friedman, Jagdish Chhabria, Diego Correa, Orli Khaimova, Richard Zheng, Stephen Haslett*

Source Code:

Assignment Objective

Build a multiple linear regression model on the Money Ball training data to predict the number of wins for the team. You can only use the variables given to you (or variables that you derive from the variables provided).

1. Data Exploration

Describe the size and the variables in the moneyball training data set. Consider that too much detail will cause a manager to lose interest while too little detail will make the manager consider that you aren't doing your job.

Dataset

Each record in the in the Money Ball training dataset 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. Below is a short description of the variables of interest in the data set:

| VARIABLE NAME | DEFINITION | THEORETICAL EFFECT |
|------------------|--|-------------------------|
| INDEX | Identification Variable (do not use) | None |
| TARGET_WINS | Number of wins | |
| 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 |
| 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 |

Figure 1: Variables of Interest

2. Data Preparation

Describe how you have transformed the data by changing the original variables or creating new variables. If you did transform the data or create new variables, discuss why you did this.

3. Build Models

Using the training data set, build at least three different multiple linear regression models, using different variables (or the same variables with different transformations). Since we have not yet covered automated variable selection methods, you should select the variables manually (unless you previously learned Forward or Stepwise selection, etc.). Since you manually selected a variable for inclusion into the model or exclusion into the model, indicate why this was done.

4. Select Models

Decide on the criteria for selecting the best multiple linear regression model. Will you select a model with slightly worse performance if it makes more sense or is more parsimonious? Discuss why you selected your model.

For the multiple linear regression model, will you use a metric such as Adjusted R^2 , RMSE, etc.? Be sure to explain how you can make inferences from the model, discuss multi-collinearity issues (if any), and discuss other relevant model output. Using the training data set, evaluate the multiple linear regression model based on (a) mean squared error, (b) R^2 , (c) F -statistic, and (d) residual plots. Make predictions using the evaluation data set.