Daniel Prusinski
Predict 412

Programming Assignment 4: Hierarchical Models

In 2011, the MLB recorded the lowest attended game with the Marlins playing the Reds, in Florida, with only 347 people in attendance, which was a result of hurricane Irene. A major source of income for Major League Baseball (MLB) is live attendance in the various team stadiums throughout the United States. Different franchises have unique strategies for selling tickets, but there are common variables that all teams face when comparing attendance.

The following exploratory data analysis (EDA) employs hierarchical and linear regression modeling methods on the bobbleheads_v002 (BHV2) dataset. All outputs, graphs, and visualizations are found within Appendix 1. Within the BHV2 dataset, there are 2,421 observations and 14 original unique variables (Output 1), of which 2/3 will be partitioned into a training set and 1/3 into a testing set (Output 2). Through the EDA, the end goal is to predict the integer variable “Attend”, which represents game attendance for each unique row. Given that the response variable is linear, the EDA will utilize modeling techniques that are appropriate for this specification.

Hierarchical models are specifically useful for data structures where data similarities adhere to organizational structures rather than a random sample. Baseball is an organized sport with non-random teams that attract non-random fans; also certain days of the week prove to attract better attendance. The natural hierarchies of such occurrences present observations that are not independent from one another and are a violation of modeling technique assumptions. Hierarchical modeling assuages the issue through creating a function for the levels (factors) and utilizing common variables to further predict based on the function.

Multiple linear regression is used as a base model to showcase the different variables as well as establish baseline metrics. Before modeling, data transformations are needed to establish variables that are factors, which will be used as hierarchies later on in the EDA. For example; variable DayofWeek has better attendance on the three weekend days as seen in Output 3, summer months are favored to Fall months (Output 4), the Dodgers, Yankees, and Phillies draw large crowds as a team and as an opponent, clear and cloudy weather are better attended than dome and rainy games (Output 5), and the promotion variable has been converted to a factor variable of which games with promotions have better attendance than no promotions (Output 6). Stepwise linear regression was used to fit the data and the model produced an adjusted RMSE of 10,024 on the training set and 9,936 on the testing set and an adjusted R-Squared of .7493.

Each factor variable was fit as an individual hierarchy model as the intercept and Outputs 8 – 13 show the details. Team was the optimum hierarchy in regard to the r-squared on out-of-sample data with .6223. The same variables were used for prediction as in the stepwise regression model, and random effects with Team and Promotion as hierarchy produced an out-of-sample model that had an r-squared of .7309 (Output 14). For the team hierarchy model, the RMSE was 6,114 on the testing data set, which is 38% better than the linear regression model. The hierarchy models are better than regular linear regression because they account for the non-random observations. If one were to analyze the residuals for the linear regression model, they would not be independent and random, which is a violation for linear regression. The hierarchical model was a better fit for the data in that it accounted for the organizational structure of the data as well as the linear variables.