Comparing NCAA Softball Performance by Division

An Executive Summary by Victoria Hannaford and Taylor Womack

Background: We procured NCAA Division I, II, and III women's softball data and performed statistical inference on it using multiple ANOVA, Kruskal-wallis test, Support Vector Machine (SVM) regression, k-fold and Monte Carlo cross validation, Nemenyi multiple comparisons, and Scheffé and Tukey pairwise comparisons to compare individual and team batting averages and see if there was anything statistically significant between divisions. We also looked at data from the NCAA Financial Database to compare the financial habits of schools by division in order to assess potential causes for the difference in the rate of injuries across divisions. (D3 > D2 > D1)

Methods: Using R, we fitted full models to predict Batting Average (BA); we also created box plots and other graphics and ran ANOVA tests to see if there were differences between each division; we used Scheffé and Tukey pairwise comparisons to examine where differences were and checked model performance using k-fold and Monte Carlo cross validation. Besides Division and BA, our models also included Games played (G), At Bats (AB), and Hits (H). Additionally, we compared the parametric ANOVA and Scheffé and Tukey pairwise comparisons to nonparametric tests including Kruskal-Wallis and Nemenyi multiple comparisons.

Results: There is a statistically significant difference between each division for team and individual BA, G, AB, and H (all p-values were < 2.2e-16). Our pairwise comparisons both showed that each division is different from the other two with p-values <2.2e-16. The SVM model for individual hits was better at predicting our model compared to the team batting average SVM model. For Monte Carlo cross validation and k-fold cross validation, we found very low RMSEs of 0.03 for the team batting average SVM model. The individual hits SVM model had an RMSE of 6.503 for the Monte Carlo cross-validation. In examining the NCAA Financial Database, we found that, in comparison to D2 programs, D1 programs spend double the money on Direct Facilities & Support for Players (17.8% vs 1.8%) and six times more on Student Athlete Meals (Non-Travel) (1.9% compared to 0.3%), while D2 programs spend 33% more than what D1 programs do on Medical Expenses and Insurance (0.9% compared to 1.2%).

Conclusion: There are significant differences by division for NCAA Softball performance. The higher batting average in D3 and D2 is likely due to having a lower quality pool of pitchers, and the higher instance of injuries in these divisions (in comparison to D1) is possibly correlated with the greater investment D1 programs put into player support. The greater incidence of injury at D2 and D3 schools is supported by the higher percentage of the budget that goes to Medical Expenses and Insurance. In the future, it would be interesting to see if there are differences between programs within a division using MANOVA or clustering.