Diamond in the rough: Data mining for predictions of student performance
Anne Alicia Kelton¹, Erika G. Offerdahl², Jeffrey Boyer²
¹Lee University, ²North Dakota State University

Introduction
SAT scores and entering GPA have been identified as significant predictive variables of student performance in introductory STEM courses. Similarly, gender has been correlated with differences in student achievement. The overarching goal of this study was to identify variables that significantly predict student achievement in an upper-level STEM course. Linear regression models and Bayesian estimations are powerful statistical tools that allow researchers to determine the predictive value of certain variables through inter- and intra-variable comparison.

Research Objectives:
1. Use stepwise regression to create a predictive model for students’ first exam scores using quantitative measures of students’ academic preparation.
2. Employ Bayesian estimation to determine if differences exist in student achievement between males and females.

Methods

Data from Entrance Survey and Assessments

Selection and Definition of Regression Variables

Stepwise Regression on Quantitative Data

Bayesian Estimation on Gender Data

The RMSE decreased from a naive baseline model of 8.58 to 8.17 when the final linear model was used to predict exam one scores using the testing data.

Findings
• Final linear model includes four variables: scientific aptitude rating for activities at the knowledge level, self-rated preparation score from prior college chemistry courses, self-reported GPA, and IMCA pre-test score.
• R² Value = .44
• GPA and IMCA pre-test scores would be expected as predictive variables.
• Chemistry and Knowledge are more likely to be specific to this model.

Future Directions
• Perform stepwise regression for other exams to see if the most predictive variables change throughout the course.
• Test the same variables from future semesters to confirm which variables are most predictive for this course.
• Attempting to generalize the variables for use from semester to semester may reduce the predictive power of the model.

References

Acknowledgements
• Thanks to Jesse Alexisnon, Andrew Calabasine, and Shannon Anderson for their input.
• Thanks to COER REU faculty, participants, and North Dakota State University.
• Thanks to the National Science Foundation for the funding for this project—Any opinions, findings, and conclusions expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.