Diamond in the rough: Data mining for predictions of student performance

Anne Alicia Kelton[1], Erika G. Offerdahl[2], Jeffrey Boyer[2]

[1]Lee University, [2]North Dakota State University

Introduction

SAT scores and entering GPA have been identified as significant predictive variables of student performance in introductory STEM courses[1]. Similarly, gender has been correlated with differences in student achievement[2]. 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

Variables	Mean	Standard Deviation
Knowledge	2.94	0.85
Comprehension	2.92	0.47
Application	2.99	0.53
Analysis	2.79	0.56
Synthesis	2.67	0.57
Evaluation	2.87	0.75
Biology	2.40	0.70
Chemistry	2.55	0.69
GPA	3.49	0.72
IMCA.PRE	53.26	16.18
EXAM.1	36.46	9.57

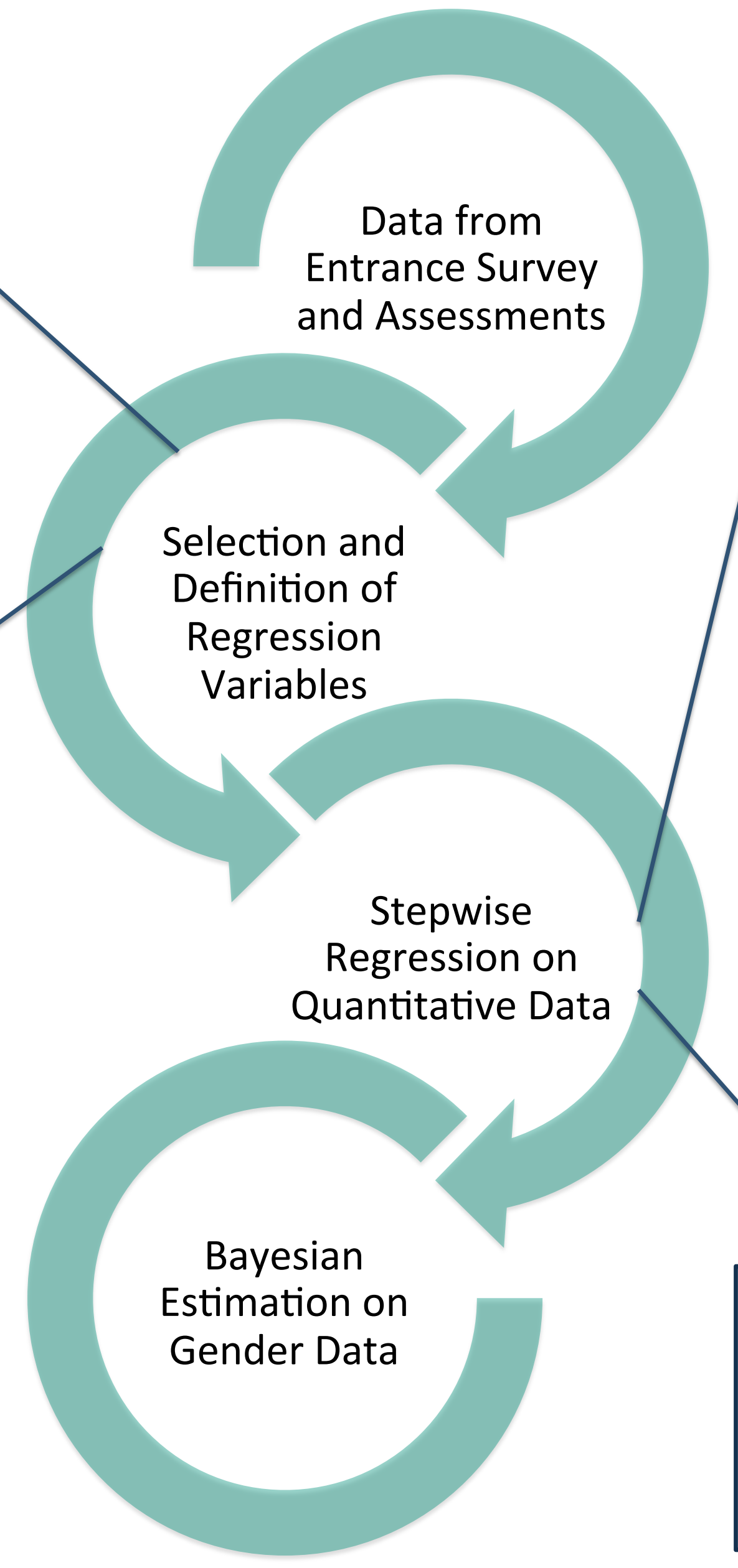
Scientific Aptitude Scores by Bloom's Level of Activities

Self-Rated Preparation

Self-Reported GPA Range

Introductory Molecular and Cell Biology Pre-Test Scores

Grades from Exam 1

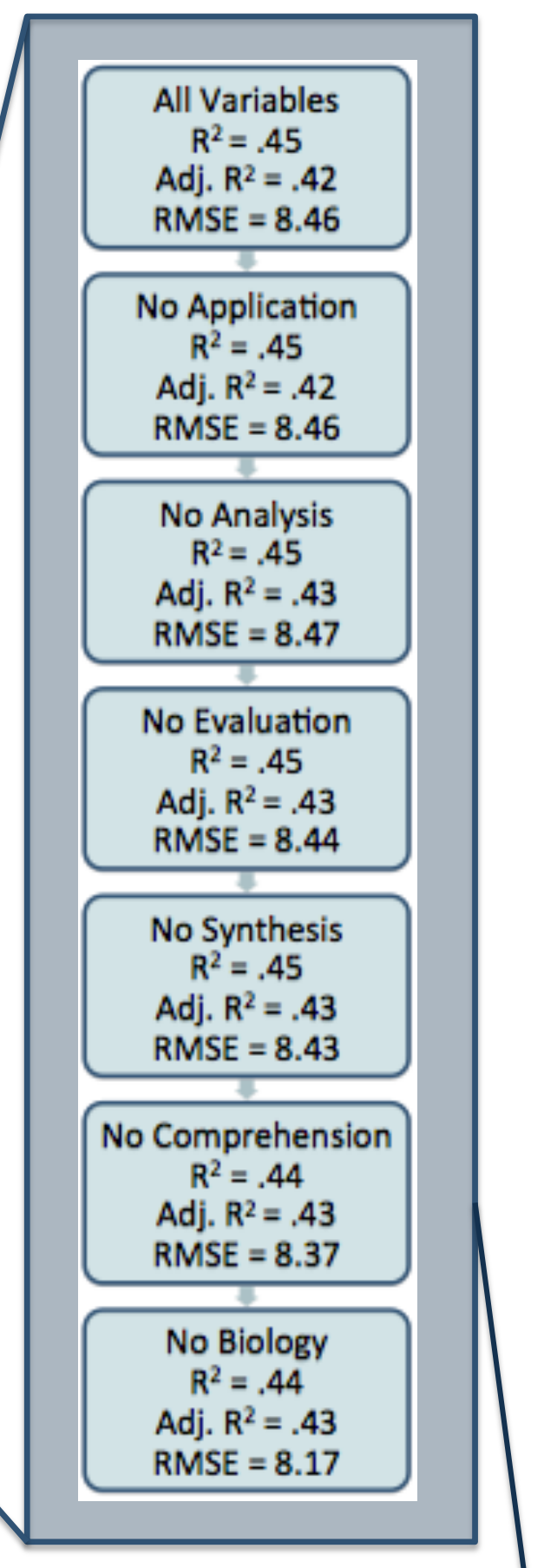


Data from Entrance Survey and Assessments

Selection and Definition of Regression Variables

Stepwise Regression on Quantitative Data

Bayesian Estimation on Gender Data



All Variables
$R^2 = .45$
Adj. $R^2 = .42$
RMSE = 8.46

No Application
$R^2 = .45$
Adj. $R^2 = .42$
RMSE = 8.46

No Analysis
$R^2 = .45$
Adj. $R^2 = .43$
RMSE = 8.47

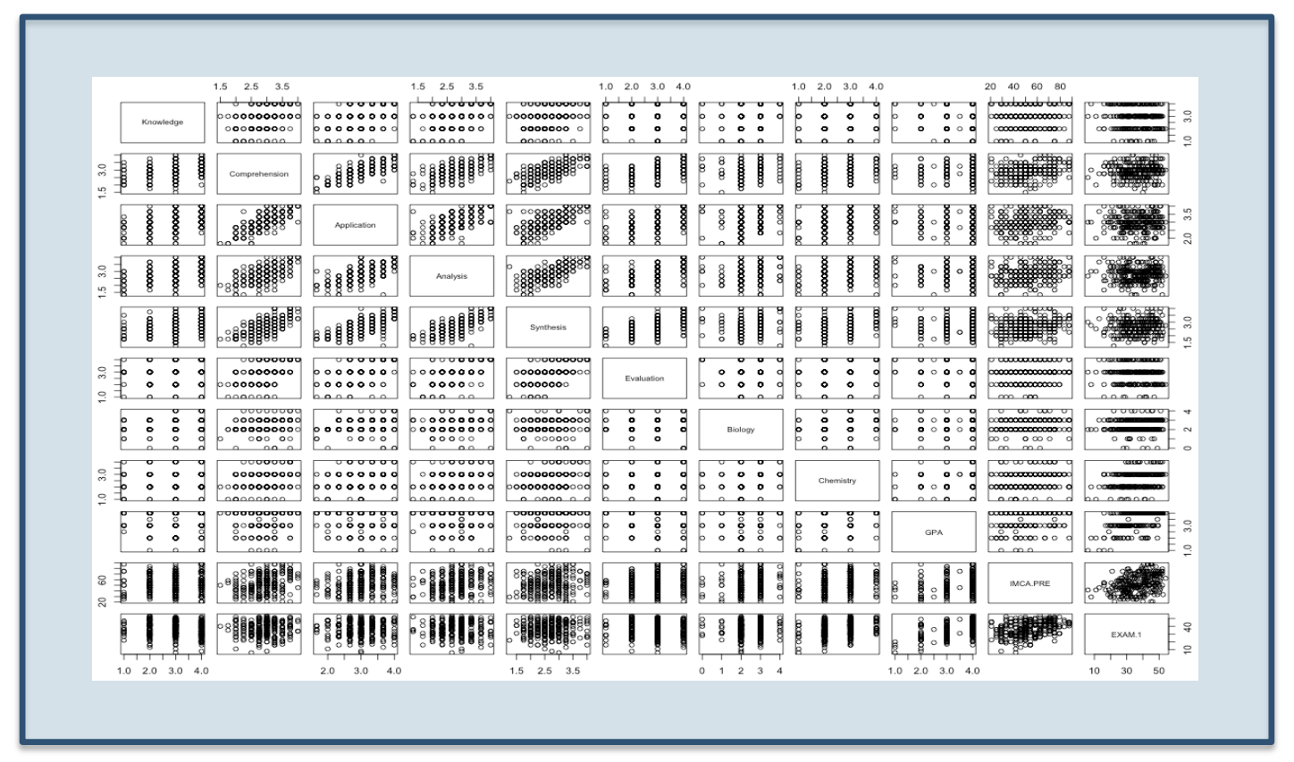
No Evaluation
$R^2 = .45$
Adj. $R^2 = .43$
RMSE = 8.44

No Synthesis
$R^2 = .45$
Adj. $R^2 = .43$
RMSE = 8.43

No Comprehension
$R^2 = .44$
Adj. $R^2 = .43$
RMSE = 8.37

No Biology
$R^2 = .44$
Adj. $R^2 = .43$
RMSE = 8.17

Linear regressions run on training data before models are used to predict and calculate RMSE from testing data.

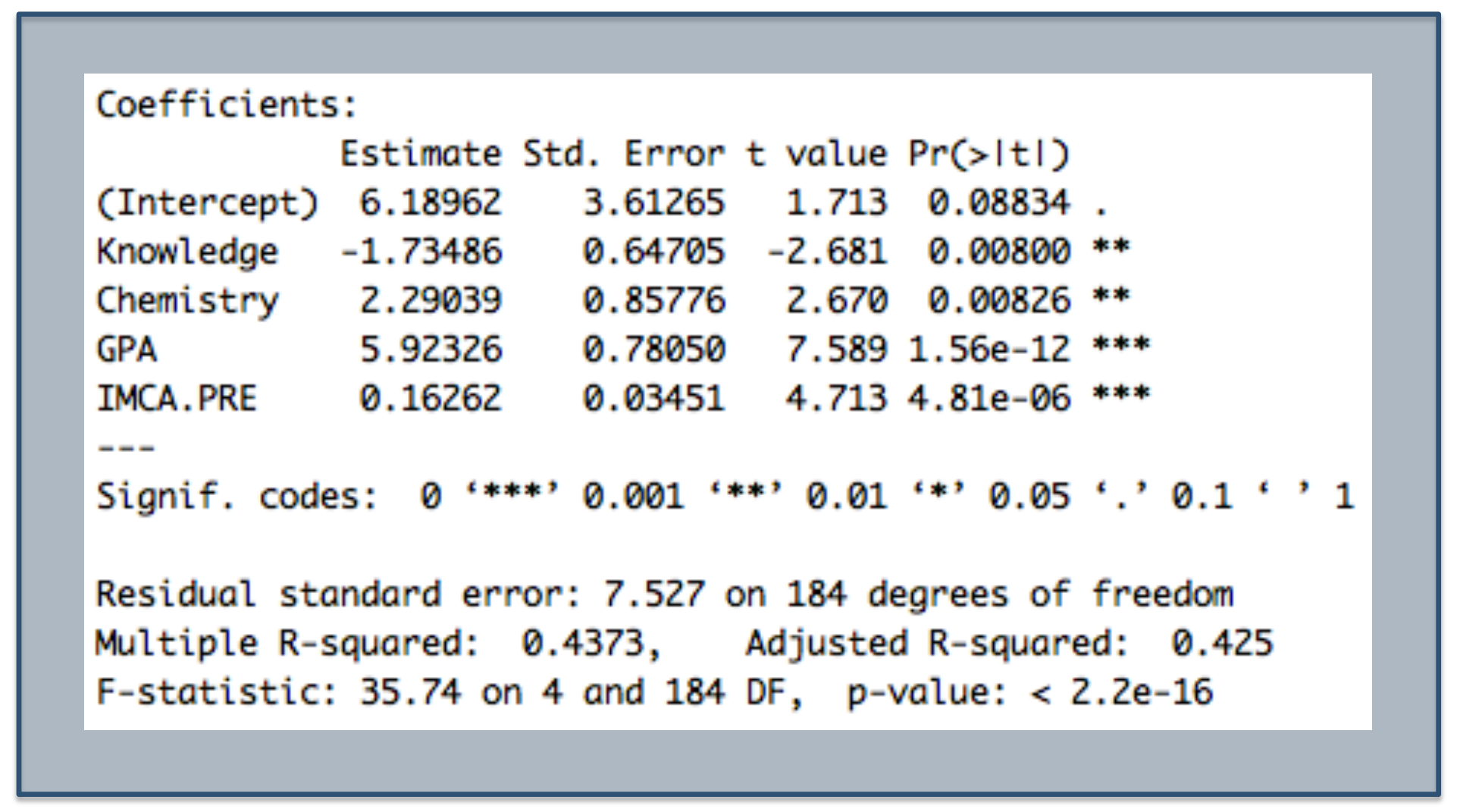


Dependent Variables	r	p	r^2
Knowledge	-0.09	0.13	0.09
Comprehension	0.07	0.26	0.00
Application	0.04	0.52	0.00
Analysis	0.02	0.69	0.00
Synthesis	0.05	0.47	0.00
Evaluation	-0.03	0.62	0.00
Biology	0.12	0.06	0.01
Chemistry	0.31	0.00	0.09
GPA	0.53	0.00	0.28
IMCA.PRE	0.36	0.00	0.13

Initial analysis of the data revealed several variables with moderate correlations; however, it did not allow for the observation of multivariate interactions.

What measures of student academic preparation can best predict a student's performance on exams?

Analysis

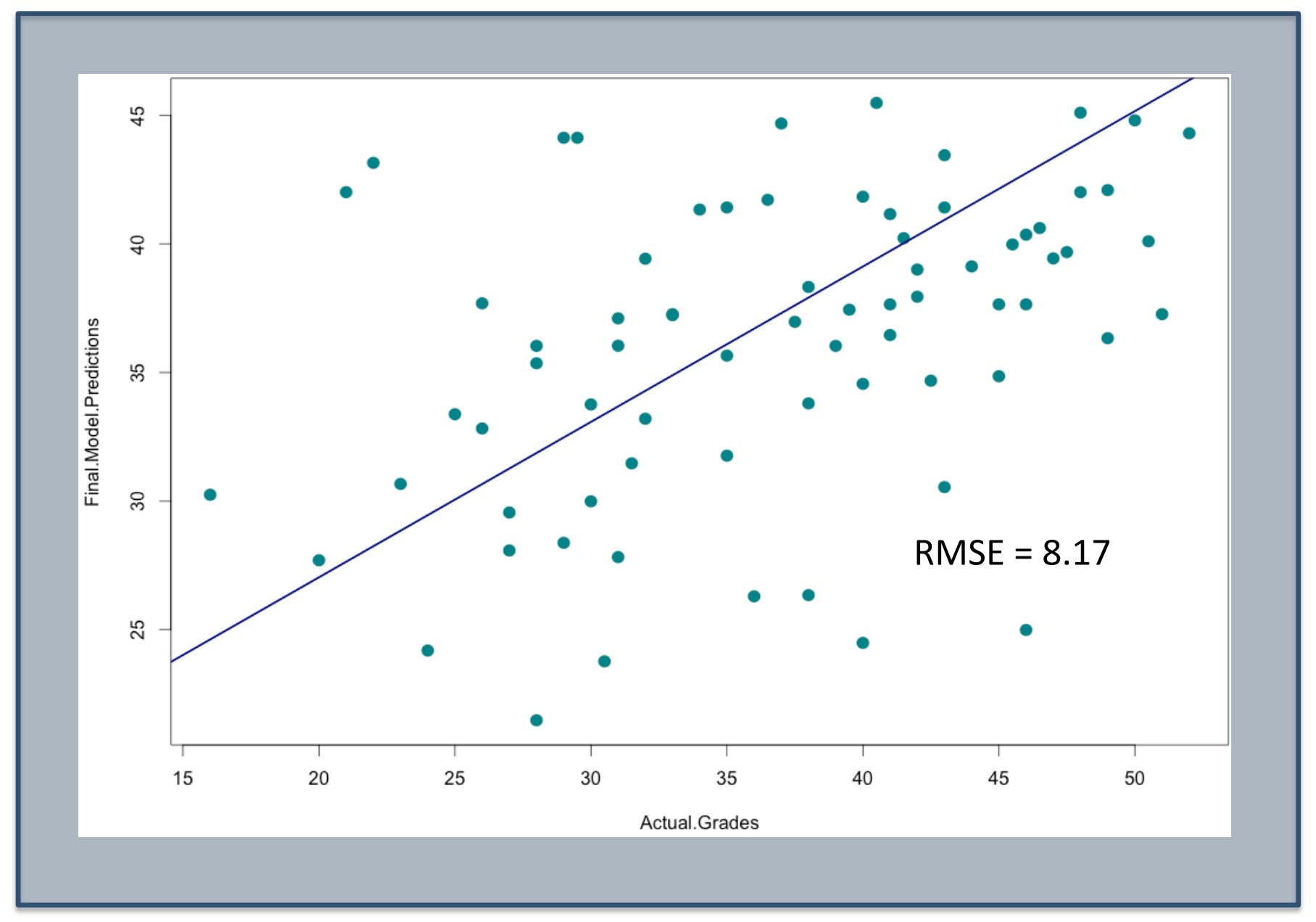


```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.18962    3.61265   1.713  0.08834 .
Knowledge   -1.73486    0.64705  -2.681  0.00800 **
Chemistry    2.29039    0.85776   2.670  0.00826 **
GPA          5.92326    0.78050   7.589 1.56e-12 ***
IMCA.PRE     0.16262    0.03451   4.713 4.81e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.527 on 184 degrees of freedom
Multiple R-squared:  0.4373,    Adjusted R-squared:  0.425
F-statistic: 35.74 on 4 and 184 DF,  p-value: < 2.2e-16
```

If the variables used had been based solely on their significance in a simple correlation matrix, the knowledge variable would not have been included in this model, reducing its predictive power.

The RMSE decreased from a naïve 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

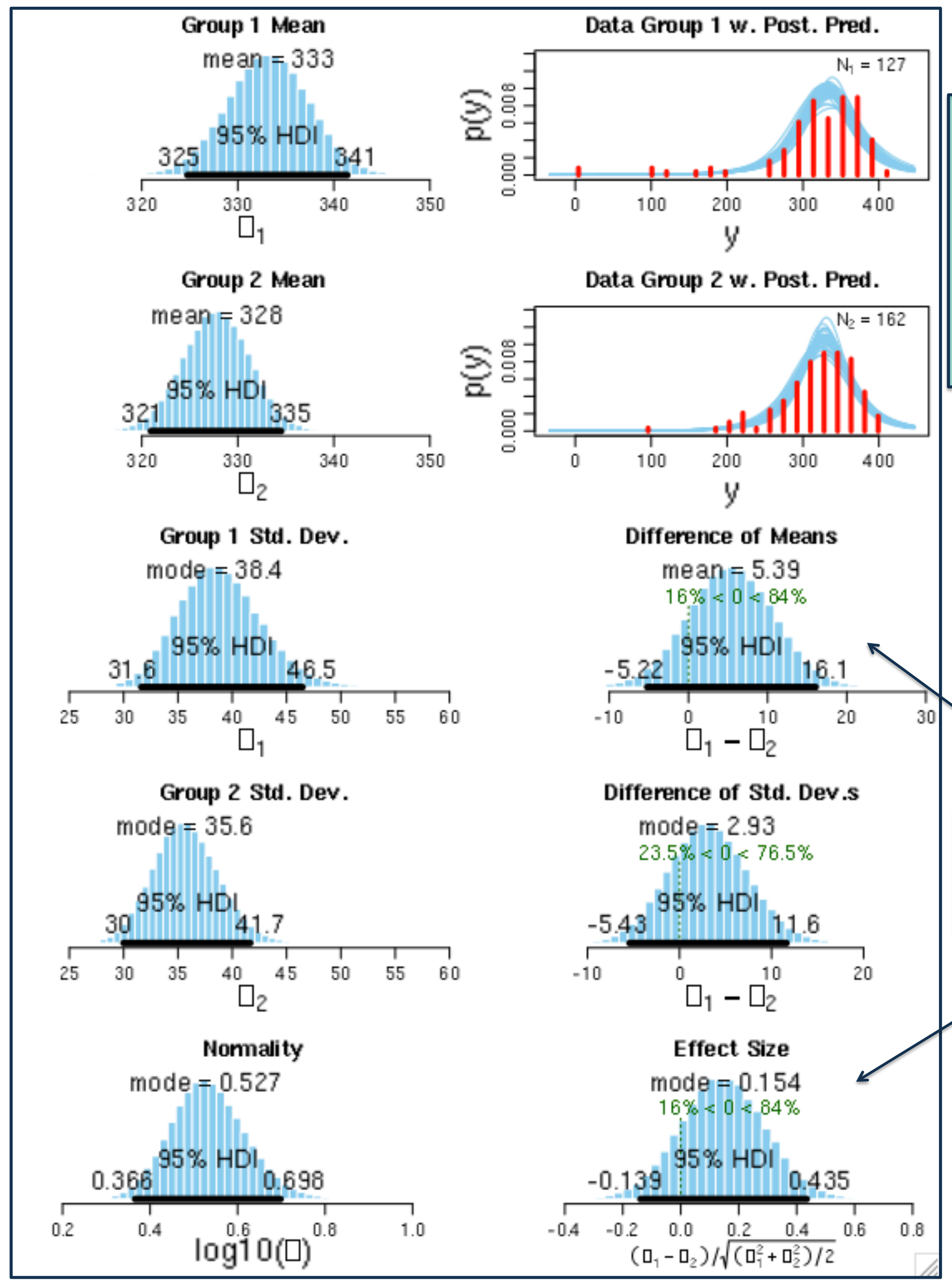
- Final linear model includes four variables:
  - scientific aptitude rating for activities at the knowledge level
  - self-rated preparation score from prior college chemistry courses
  - self-reported GPA
  - IMCA pre-test score
- R^2 Value = .44
- GPA and IMCA pre-test scores would be expected as 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[3].

Are there differences in achievement between males and females in the course?

Analysis



Distribution of credible values superimposed over a histogram of the sample values.

Distribution of credible values

Group 1 = 127 male students
Group 2 = 162 female students

Findings

- Zero is a credible difference of means between males' and females' overall course performance.
- Zero is a credible effect size.
- There is no credible difference between how male and female students perform overall.

Future Directions

- Repeat the test on data from future semesters.
- Use Bayesian estimation to test for differences between the overall performances of different majors.
- Specifically, do the pharmacy majors that makeup a majority of the population perform differently than other majors represented in the course?

References

1. Theobald, R., & Freeman, S. (2014). Is It the Intervention or the Students? Using Linear Regression to Control for Student Characteristics in Undergraduate STEM Education Research. *CBE-Life Sciences Education*, *13*(1), 41-48.
2. Miyake, A., Kost-Smith, L. E., Finkelstein, N. D., Pollock, S. J., Cohen, G. L., & Ito, T. A. (2010). Reducing the gender achievement gap in college science: A classroom study of values affirmation. *Science*, *330*(6008), 1234-1237.
3. Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers & Education*, *54*(2), 588-599.
4. Morris, L. V., Finnegan, C., & Wu, S. S. (2005). Tracking student behavior, persistence, and achievement in online courses. *The Internet and Higher Education*, *8*(3), 221-231.

Acknowledgements

- Thanks to Jessie Arneson, Andrew Calascione, and Shannon Anderson for their input.
- Thanks to CiDER 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.

NSF-DUE 1156974