IOWA STATE UNIVERSITY
Department of Animal Science

2020 ASAS-CSAS-WSAS
Virtual Annual Meeting & Trade Show
Late-Breaking Abstract
July 23, 2020

Nick V. L. Serão*, Amy A. Petry, Leticia P. Sanglard

*serao@iastate.edu

Self-reported statistical training of graduate students associated with confidence in performing statistical analyses

Introduction

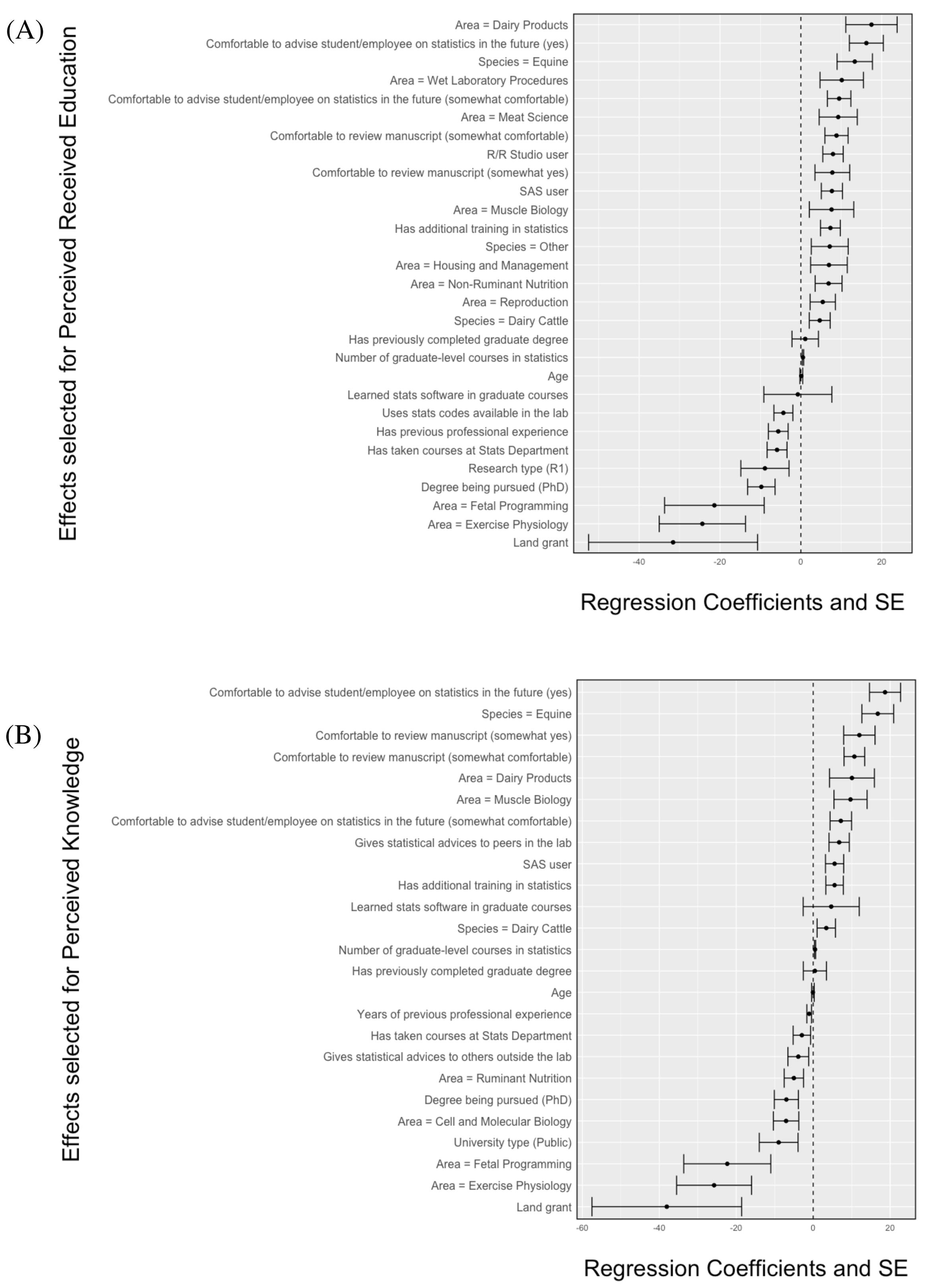
- Statistical analysis of data is one of the most important aspects of graduate education in Animal Sciences;
- The development of high-throughput phenotyping and *Big Data* technologies requires graduate to be skilled in management and analysis of large amounts of data;
- Therefore, the Animal Science graduate students of today must be highly exposed and trained in statistics to address the academic and industry needs of the future;
- **The objective of this study was to identify the current statistical competencies associated with graduate students' perceive received education, knowledge, and confidence to perform statistical analyses.**

Material and Methods

- Survey data from students enrolled in Animal Science-related graduate programs (N=416; 153 M.S. and 263 Ph.D. students), representing 43 U.S. universities, were used;
- Four groups of questions were used in the survey:
  - 1) Demographics and overall training (*e.g.* age, years of graduate education, professional goal, degrees completed, etc);
  - 2) Perceived Received Education (**PRE**) on 30 topics (*e.g.* Computer Coding, ANOVA, Machine Learning, Observational Studies, etc.);
  - 3) Perceived Knowledge (**PK**) acquired on the same 30 topics as above;
  - 4) Confidence in performing statistical analysis (**CPSA**) on 31 topics (*e.g.* Designing Experiments, Data Management, Mixed Model Analysis, etc);
- Answers (i.e. scores) followed a 6-point scale: 0 to 5, representing no to high quality of education received (PRE), knowledge (PK), and confidence (CPSA);
- Quality control: within each of the groups of questions (PRE, PK, and CPSA), individuals with more than 10% of missing answers and/or with all answers being the same were removed;
- The remaining missing answers were imputed within each group of questions using a bootstrap-based EM algorithm (Honaker and King, 2010). A total of 10 bootstraps imputed runs were generated and the rounded average answer was used in the final dataset;
- Answers followed a 6-point scale: 0 to 5, representing no to high quality of education received (PRE), knowledge (PK), and confidence (CPSA);
- Quality control: within each of the groups of questions (PRE, PK, and CPSA), individuals with more than 10% of missing answers and/or with all answers being the same were removed;
- The remaining missing answers were imputed within each group of questions using a bootstrap-based EM algorithm (Honaker and King, 2010). A total of 10 bootstraps imputed runs were generated and the rounded average answer was used in the final dataset;
- Overall scores for each participant were calculated as the sum across scores within each of the groups of questions: PRE, PK, and CPSA;
- Statistical Analysis:
  - Associations between demographic and overall training answers with PRE and PK overall scores using backward selection based on AIC;
  - Spearman's correlation and Ward's hierarchical clustering between PRE and PK answers;
  - Cluster analysis of CPSA scores using DIvisive ANAlysis (DIANA) based on Gower's dissimilarity;
  - Associations between demographic, overall training answers, and PRE or PK scores on CPSA overall scores using backward selection based on AIC. Analyses were performed separately using either PRE or PK scores.

Conclusions

- Demographics and overall training had limited impact on explaining the variation of how graduate students perceive their received education and knowledge;
- The perceived received education and perceived knowledge on specific topics were only moderately correlated;
- Their comfort in performing statistical analyses seem to be broadly divided into two groups: common and complex statistical methods;
- Most topics of perceived received education and perceived knowledge were positively associated with their overall comfort in performing statistical analyses, indicating that a better perceived training increases their overall knowledge;
- Additional studies are needed to objective test graduate students on statistics.



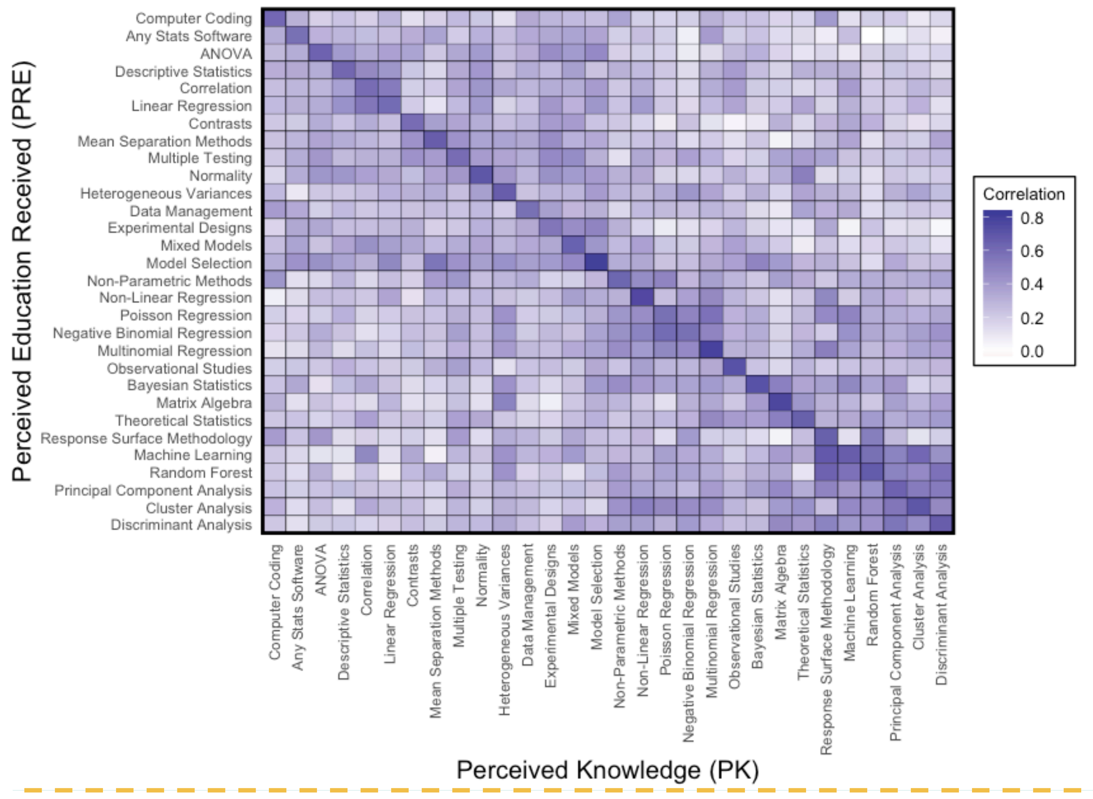
Effect of demographic and overall training answers on Perceived Received Education (PRE; **A**) and Perceived Knowledge (PK; **B**)

A total of 29 (R^2 = 0.354) and 25 (R^2 = 0.396) variables were selected for PRE and PK, respectively;

Unless specified within parenthesis, estimates of categorical effects represent the effect of the answer "Yes" in comparison to "No";

For both analyses, participants from Land-grant institutions showed lower PRE and PK than those from non-Land-grant institutions;

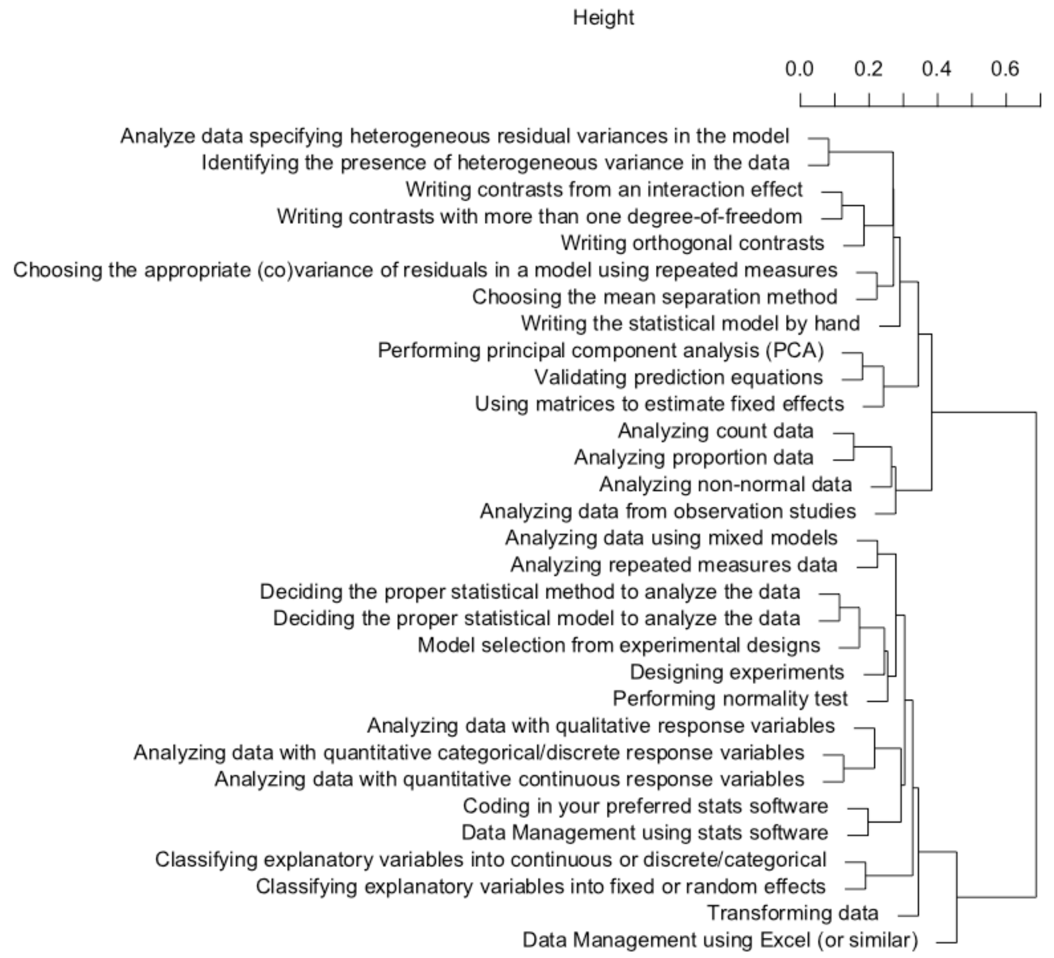
In contrast, in general, participants that are comfortable to give statistical advices in their future career shower higher PRE and



Marginal Spearman's correlation between PRE and PK answers

Correlation coefficients across the diagonal ranged from 0.53 between the PRE and PK of *Any Stats Software* to 0.77 between the PRE and PK of *All Matrix Algebra*;

Off-diagonal correlation coefficients ranged from 0.04 between PRE on *Any Stats Software* and PK on *Machine Learning* to 0.65 between PRE on *Machine Learning* and PK on *Random Forest*;

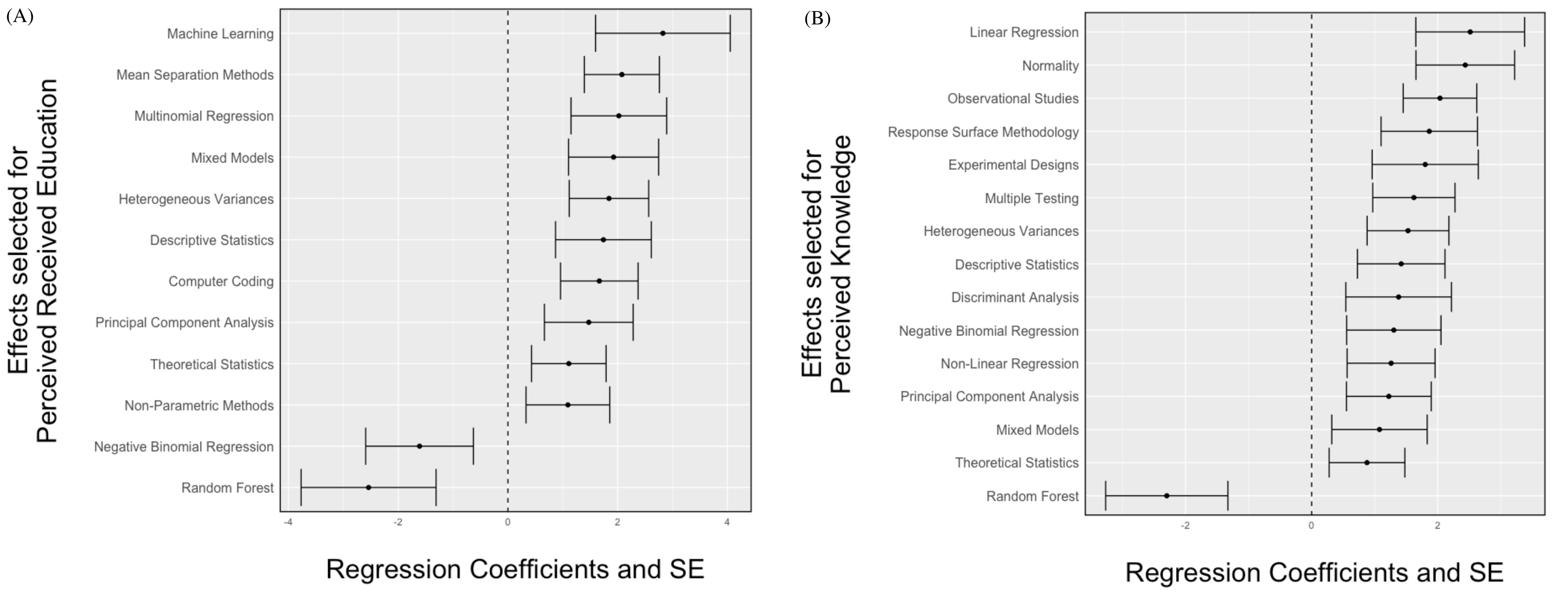


Dendrogram of clustered CPSA scores

Divisive cluster analysis based on Grower's dissimilarity matrix for the 31 CPSA topics;

Two major clusters were formed, with 15 and 16 topics in each;

In general, the top cluster included the use of complex statistical methods, such as analysis of count and non-normal data, machine learning algorithm, etc. In contrast, topics included in the bottom cluster, such as classification of variables, data management, etc.



Effect of selected Perceived Received Education (PRE; **A**) and Perceived Knowledge (PK; **B**) on overall Confidence in Performing Statistical Analysis (CPSA scores)

Figures above show only the PRE (A) and PK (B) topics associated with CPSA – the demographic and overall training variables were excluded to enhance visualization;

A total of 64 (R^2 = 0.727) and 60 (R^2 = 0.727) topics were selected in the model for the PRE and PK analyses, respectively. Of these, 12 PRE and 15 PK topics were selected, respective, with 7 in common to both analyses;

In general, topics had a positive effect on CPSA: the higher was the score on each topic, a greater CPSA is expected;

The topic of *Random Forest* had a negative impact on CPSA for both analyses. *Negative Binomial Regression* had a negative effect on CPSA for the PRE analysis and a positive effect for the PK analysis;

Across both analyses, *Machine Learning* had the greatest effect on CPSA;