An Application of Exploratory Data Mining in Project TALENT

Ross Jacobucci[1] John J. McArdle[1] John J. Prindle[2]

University of Southern California, Department of Psychology[1];
University of Southern California, Department of Social Work[2]

jacobucc@usc.edu

September 19, 2015

Overview of Project TALENT

- Research project directed Dr. John C. Flanagan and conducted by the American Institutes for Research (AIR) and the University of Pittsburgh.
- The baseline data collection occurred in 1960 with more than N $\approx$ 377,000 ($3.77x10^5$) participants.
- Overall goals of Project TALENT (PT) were to characterize the abilities, interests and aspirations of U.S. high school students and to see how these predicted educational and occupational outcomes. Students were tested across 2 full days or 4 half-days.
- Data are available through the ICPSR website (#33341).

Factor Structure of Intelligence in PT

PT offers a unique opportunity to examine the factor structure of intelligence due to both its large sample size, but maybe more importantly, variety of cognitive ability questions. The purpose of this study was not re-examine and possible derive a "best" fitting model, as this was the motivation behind a number of previous studies (e.g. Major, Johnson, & Deary, 2012).

The factor structures used in this study were based on the work of McArdle et al. (2015) which took a confirmatory approach to testing a number of a priori specified models using a subset of the cognitive items.

Exploratory Data Mining

The past 10 years have seen an emergence of Exploratory Data Mining (EDM; McArdle & Ritschard, 2013) applications in the social and behavioral sciences.

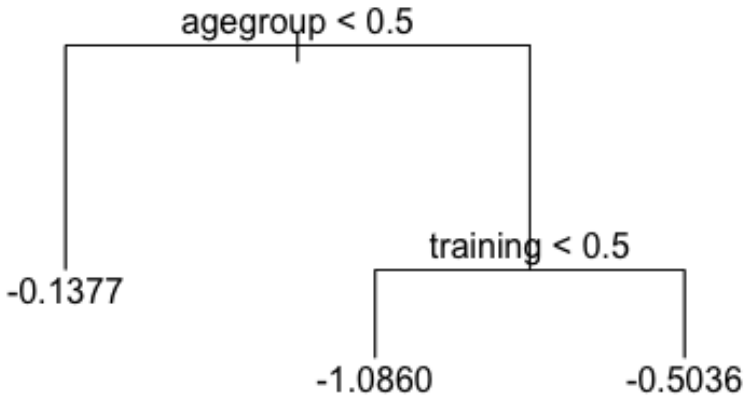
The emergence of EDM has also increased the necessity of delineating between "confirmatory" and "exploratory" methods.

Tree based methods are one of the most widely applied EDM methods. These include Decision Trees (e.g. CART; Breiman, Friedman, Olshen, & Stone, 1984), Random Forests, and Gradient Boosting Machines.

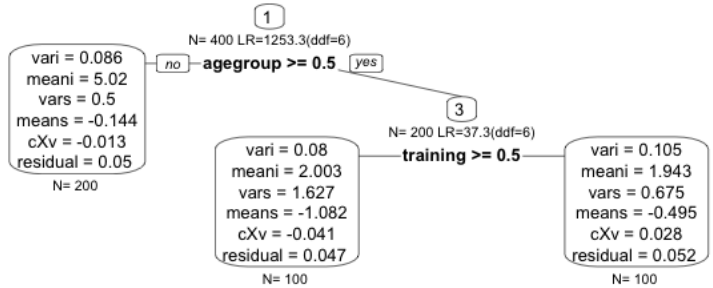
Just because we have a fancy, powerful, new tool doesn't mean we need to use it.
A generalization of Decision Trees include Structural Equation Model Trees (SEM Trees; Brandmaier, von Oertzen, McArdle, & Lindenberger, 2013).

SEM Trees

Example From CART



Example From SEM Trees



Instead of creating binary splits on covariates in relation to a singular outcome as in Decision Trees, SEM Trees split based on improvements in SEM model fit, comparing the likelihood ratio between the two group split and no split (no groups).

Can be thought of as doing an "automated" search for multiple group models.

Implemented as the *semtree* package in R. Downloadable from:

http://brandmaier.de/semtree/

Our Study

Based on the a priori specification of 4 different factor models of intelligence, using 22 cognitive items on a subset of 10,000 cases, what could we learn from using SEM Trees to search for covariates that showed a relationship to the factor structure?

In each factor model, only the factor means, variances, and covariance were allowed to vary across groups, thus imposing strict factorial invariance.

Of the 2,105 original variables in the base year dataset, "only" 1,191 covariates were used, after eliminating other cognitive variables and variables with low response rate etc...

Looking for *Important*, not just *Significant* predictors

What did we think we would find?

- Parental Occupation and SES
- Gender Differences
- Race/Ethnicity
- School Interests/Activities/Motivation
- Type of School Attended
- **Post-Hoc** Any cognitive items that we missed in screening

Items

Cognitive Items

- Memory for Words (I=24)
- Memory for Sentences (I=12)
- Arithmetic Reasoning (I=16)
- Introductory Calculations (I=24)
- Advanced Calculations (I=14)
- five English Total tests (I=100)
- Abstract Reasoning (I=15)
- Mechanical Reasoning (I=20)
- Disguised Words (I=30)
- Creativity (I=20)
- Clerical Checking (I=74)
- Visual 2D (I=24)
- Reading Comprehension (I=48)
- Visual 3D (I=16)
- Word Functions (I=24)
- Table Reading (I=75)
- Object Inspection (I=40)

Non-Cognitive Items

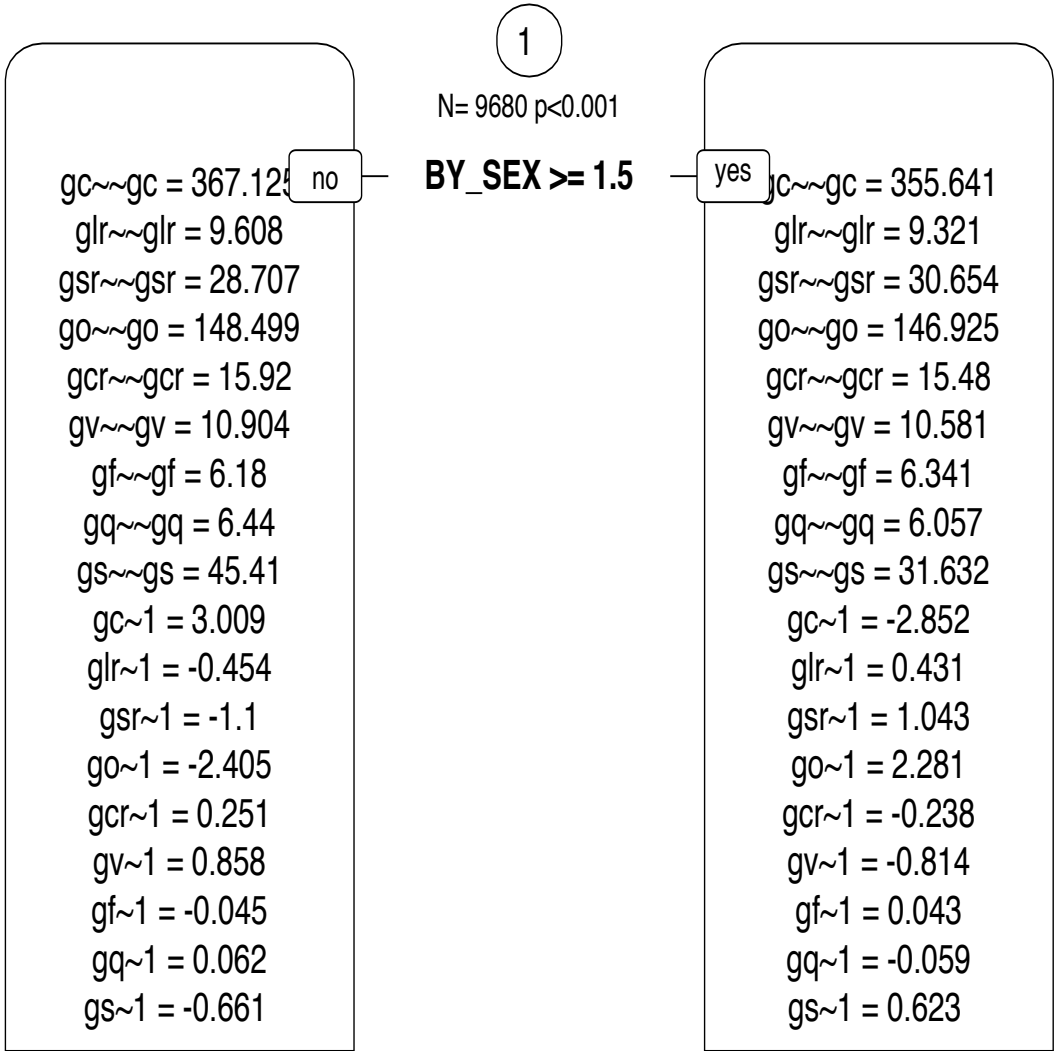
- Background information
- 17 scales of interest inventory
- Information about school attended filled out by Guidance Counselor
- 10 scales of student activities (temperament)

Factor Structure of Intelligence Tested

Factor Models included:

- one first-order factor model
- a two-factor Dual Process Model
  - Type 1 vs. Type 2; (Kahneman, 2011; Stanovich, 2011)
- 8 first-order factors
  - Gf-Gc w/o 2nd Order factor (Horn, 1966; Woodcock, 1990, 1993)
- Orthogonal bi-factor model with 8 specific factors

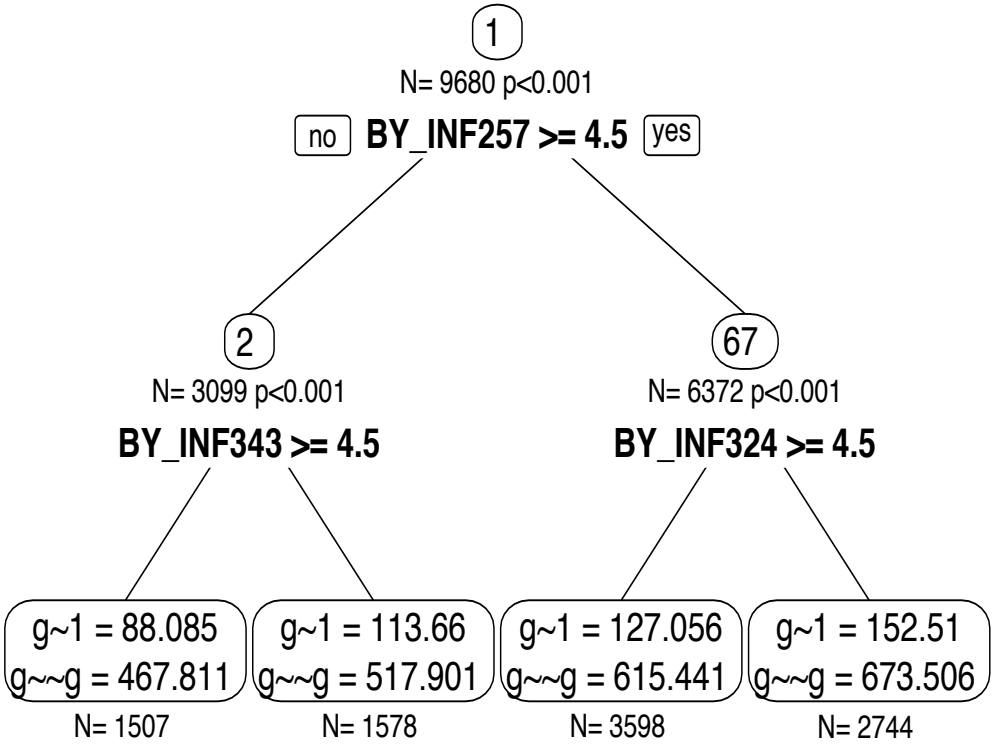
Results



Male = 1, Female = 2

Results

Example Model: First 2 levels from One-Factor Model



Important Covariates

- **BY_INF257**: An allergy is an abnormal 1. Craving, 2. Personality, 3. Fear, 4. Physique, **5. Sensitivity.**
- **BY_INF343**: "moving the two oars at once is better than alternating them because it is" 1. Safer, 2. Easier to Learn, 3. Better Exercise 4. Less Likely to Splash **5. More Efficient**
- **BY_INF324**: "A license is not needed in order to be a" 1. Ham Radio Operator, 2. Truck Driver, 3. Physician 4. Private Pilot **5. Physicist**
- **BY_INF323**: "A newspaper editorial usually contains" 1. An Advertisement 2. A Photograph 3. Facts Without Interpretation **4. An Opinion or Interpretation** 5. A Fictional Treatment
- **BY_SIB304**: "Greatest Amount of Education Expected" split between advanced college degree and less schooling
- **BY_D802** HS curriculum: academic versus other.

What Weren't As Important

- Nothing from:
  - Parent's Education
  - Family Income
  - Nothing regarding intellectual stimulation in home
  - Any other family background characteristic
- Only Race/Ethnicity predictor was African American composition of schools
  - Differences once makeup of school was above 30-40%
- Student Interests Played Small Role
  - Potential College Major
  - Interest in School / Motivation / Attention

Conclusions

- SEM Trees took a long time to run...
  - 2-3 weeks on average
- SEM Trees allows you to look for differences with respect to individual parameters
- With so many covariates, able to answer what is of overall importance, not just significant
  - SEM Forests in the future
- More suprising was what **wasn't** important.
- Like with all EDM, best to cross-validate.

The End

contact: jacobucc@usc.edu
For more info, scripts, etc...

slides available at:
www.github.com/Rjacobucci/ISIR_talk2015