Self-Assessment 3

Signal Data Science

We'll be having another self-assessment. As before,

- Type your answers in a new R script file with comments indicating where the answer to each question begins.
- When you finish, email `signaldatascience@gmail.com` with your R script attached along with the amount of time you spent on the self assessment.
- Work individually. You can however consult R documentation, look at old assignments, use the Internet, etc., but don't copy and paste code verbatim.
- Make your code as clear, compact, and efficient as possible. Use everything that you've learned! **Please comment and organize your code so we can easily tell how parts of your R script correspond to specific problems.**

The results that you'll be obtaining are interesting – have fun with it! This self assessment is long and will probably extend beyond lunch. We'll break for lunch at 12:30 PM as usual. After the self-assessment, you'll resume work on the recommender systems assignment.

National Election Study

In the `nat-elections` dataset folder, there is a cleaned version of data from the 1992 National Election Study, with demographic information about 1771 US citizens and how they voted in the 1992 US Presidential election as `nes_cleaned_1992.csv`. A glossary describing the variables and giving summary statistics for the original dataset is available in `nes-glossary.txt`.

- Expand the factors into dummy variables using `dummy.data.frame()` from the `dummies` R package and remove one dummy variable for each factor to avoid the multicollinearity dummy variable trap.
- Restrict to those people who voted, and cast their vote for either the Republican or Democrat candidate (George H. W. Bush and Bill Clinton,

respectively). Use `glm()` to model the probability of a voter casting a vote for the Republican candidate with logistic regression.

- Order the features by size in order of decreasing magnitude to determine the most predictive features. Evaluate the quality of the model by computing the area under the ROC using the appropriate function from the `pROC` package. (No need for cross validation here – the amount of overfitting is negligible. How can you tell this ahead of time?)
- Use your model to generate predictions for how those people in the study who didn't vote would have voted. What does your model predict the percent who would have supported Clinton to be? How does this percentage compare with the percentage of voters who actually voted for Clinton?

National Merit Twin Study

The `nmsqt-twin` dataset contains data from the 1962 National Merit Twin Study as `NMSQT.csv`, with data on 752 twin pairs. The features that I've included are:

- ID: The unique identifier of a twin pair. There are two entries with each ID, corresponding to two members of a twin pair.
- ZYG: A categorical variable specifying whether a twin is a member of a pair of identical twins or fraternal twins.
- NMSQT: The National Merit test scores of the participant.
- V11093 through V11572, answers to 478 questions from the 1956 California Psychological Inventory (first edition). "Yes" answers to the questions are coded 1 and "no" answers are encoded 0.

Also included is a codebook giving the text of the questions from the California Psychological Inventory as `NMSQTcodebook.csv`.

- Use the `caret` package to determine the best values of `alpha` and `lambda` for a regularized elastic net regression model for National Merit test score in terms of answers to the questions from the California Psychological Inventory. (Refer to the material from the assignment which discusses the `caret` package as needed.)

  What percent of the variance in National Merit test scores does the best model explain?

  Take the coefficients of the best model and join them to the question text in `NMSQTcodebook.csv`. Order the coefficients in order of decreasing magnitude to see which questions are most predictive.
- Do principal component analysis on the CPI questions using `prcomp()` and plot the standard deviations of the principal components in decreasing

order. Using the resulting scree plot, make an educated guess for the number of factors to use in factor analysis on the questions.

- Do oblique factor analysis on the questions with the number of factors determined above using `fa()` from the `psych` package. Bind the factor loadings to the question names. For each factor, order the loadings in order of decreasing magnitude to interpret the factor. Give each factor an appropriate name based on the questions with loadings of large magnitude. Convert the factor scores into a data frame with these names. (Refer to the factor analysis assignment as appropriate.)
- Scale the factors so that the unites are standard deviations. Aggregate the factors by gender using `aggregate()` to pick up on gender differences.
- For a given variable, The percent variance explained by additive genetic effects is

$$2 \times (r_{\text{MZ}} - r_{\text{DZ}})$$

where r_{MZ} and r_{DZ} are the correlations between corresponding members of identical twin pairs and fraternal twin pairs respectively.

For each factor and for the variable `NMSQT`, compute the percent variance explained by additive genetic effects. Interpret the results.