Logistic Regression: Speed Dating

Signal Data Science

We'll introduce logistic regression by returning to the Columbia speed dating dataset.

- Load `speeddating-aggregated.csv` in the `speed-dating` dataset. The dataset is an aggregated form of the full speed dating dataset; you've worked with a simplified form of this dataset before (with fewer variables). Refer to the documentation in `speeddating-documentation.txt` for a description of the new variables.

You can run logistic regression with `glm()`. It can be used in the same fashion as `lm()`, except for logistic regression you must pass in the additional parameter `family="binomial"`. Additionally, the column representing the binary class which you want to predict must either be (1) a numeric column taking on values 0 and 1 or (2) a factor.

The `pROC` package provides a function, `roc()`, which plots the receiver operating characteristic (ROC) curve given the results of a logistic regression fit. In addition, `auc()` can be called on the output of `roc()` to calculate the area under the ROC curve. Note that `roc()` accepts *probabilities* as inputs, but the predictions made with a logistic regression model will be in the form of *log-odds ratios*, which must be converted into probabilities with

$$P = \frac{\exp L}{1 + \exp L}$$

where L is a log-odds ratio and P is the corresponding probability.

When working through the following questions, examine and interpret the coefficients of each logistic regression model. In addition, examine the area under the ROC curve as well as the shape of the ROC curve itself.

- Predict gender in terms of the 17 self-rated activity participation variables.
- Restrict to the subset of participants who indicated career code 2 (academia) or 7 (business / finance). Use logistic regression to differentiate between the two.

- Restrict to the subset of participants who indicated being Caucasian (race = 2) or Asian (race = 4). Use logistic regression to differentiate between the two.