SimData: the data simulation game

Robert Horton

March 28, 2015

SimData is an informal contest between a data engineer and a data analyst. The engineer's goal is to generate a data set with an outcome that can be predicted accurately using a linear model, but to make the analysis process challenging by obfuscating the relationship between inputs and outcome.

Sportsmanship is key. The goal is not to make a problem that is impossible to solve, but rather to create a problem that can be solved by creative application of standard analytical approaches.

The Story

Your data must reflect a health-related scenario where the outcome can be predicted from the inputs (or some transformation of the inputs). Preferably you should provide references to articles from the biomedical or health care literature which describe similar relationships. We want the application scenarios to be fairly realistic, but the datasets should generally be simpler and cleaner than real life (once the engineered obfuscation has been sorted out).

The Data Set

Tidyness

Analysis will be done on a single table of tidy data. You may supply "raw" data in an untidy format that requires aggregation, transformation or other manipulation prior to analysis. The following criteria apply to the data set once it has been tidied.

Outcome

Your outcome variable can be continuous (e.g., blood pressure) or binomial (e.g., cancer vs. normal; whether or not the patient survived). Do not have a multinomial outcome (better | worse | no change).

Inputs

- Input variables can be continuous or categorical.
- Include at least some inputs that have no relationship to the outcome (distractors).
- Try to include some inputs that have weak relationships with the outcome, but do not add any information on top of the more important predictor variables in the dataset.
- The total number of input variables should not be less than 5 nor more than 20.

Observations

- The number of observations must be sufficient for both training and testing.
  (NEED STATISTICAL CRITERIA HERE)

Examples

Search PubMed for “linear regression” or “logistic regression.” You should filter the article for those that have “Free full text” (under “Text availability” in the left column of the results page), since these will make more widely accessible references.

Types of Obfuscation

Noise

- You can add noise to the inputs, the outcome, or both. The noisy data must still be amenable to analysis; generally, larger data sets are can compensate for noise.
- All random noise should be from the random normal distribution, and should be additive.

Categorization level

A simple relationship between category and outcome can be obscured by providing a different level of categorical detail. For example:

- An outcome (blood alcohol concentration) might be higher on particular days of the week (Saturday), but you supply the analyst with a date.
- Trauma admissions from automobile accidents might be higher for people driving sports cars; you supply the analyst with make and model of vehicle, and they need to look up which are sports cars.
- Diagnosis codes (ICD9) might be provided at a very detailed level of specificity, but the outcome is determined by a more general diagnosis.
- You can go in the other direction (providing too-general categories) by using interactions between categories (young women in certain geographical regions have higher rates of anorexia).

Transformation

- The outcome (blood pressure) has a linear relationship on BMI, but you provide height and weight instead.
- The outcome is linearly related to the logarithm of household income; you provide the income.

Invisible Explanatory Factors

- The outcome is linearly related to the subject’s ability to interpret spatial information. You provide scores from a battery of tests, some of which correlate with spatial ability, but you do not provide the spatial ability value itself.

Multicollinearity

- Some variables are highly correlated with one another (both height and arm length are provided).

Aggregation

- The outcome is related to sodium in the diet; you provide the diet (and a way to compute sodium).
- Skin cancer is related to total sun exposure; you provide activity logs (and a way to compute sun exposure)

Sportsmanship

- Remember, this is a *friendly* contest.
- We are using this as a way to review and study analytical approaches. There should be a clear path to the answer using data manipulation, feature engineering and data analysis approaches we have learned (or will learn) in class.
- You need to be able to do the analysis on your own simulated dataset yourself.
- What goes around comes around. You will need to analyze someone else's dataset.

Regression Model

Your data will be analyzed with ordinary least squares linear regression (using `lm`) if the outcome is continuous, or with logistic regression (using `glm` with `family = "binomial"`) if the outcome is binomial.

References

- ICD-9 tobacco use codes are effective identifiers of smoking status
- Confounding and Collinearity in Multivariate Logistic Regression
- Problems of correlations between explanatory variables in multiple regression analyses in the dental literature