Clinical Statistics Methods Forum Data Challenge

Eric Polley & Prabin Thapa

October 18th, 2016

Introduction

The 2016 Clinical Statistics Methods Forum Data Challenge
The goal is to construct a predictor of therapeutic drug dose given baseline clinical measurements. Participants given a training data set with 3,000 patients and will be evaluated on a blinded validation data set with 1,722 patients

FAQ

- Teams can submit as many predictions on the validation data set as they want, but I'll randomly select only one for evaluation
- Two primary loss functions:

$$L_A(\hat{y}, y) = (\hat{y} - y)^2 \tag{1}$$

  And

$$L_B(\hat{y}, y) = \mathrm{I}(|\hat{y} - y| > 16) \tag{2}$$

- The equation $L_B(x, y)$ is the indicator the predicted dose is more than 1 standard deviations away from the truth
- These notes will be available at `https://github.com/ecpolley/CSMF_Data_Challenge`

Plan for November

- Please submit predictions on the validation data prior to Nov. 14th, 5:00PM (central)
- Meet on Nov. 15th, 2:00PM
- Each team will have the opportunity to describe their analytic approach with 1-2 slides
- Briefly describe how you processed the data and your approach for estimating the predictor
- I will then summarize the different methods and reveal the validation set performance with discussion

How To Get Data

```
# link to data on GitHub page if not available
if(file.exists("TRAIN.CSV")) {
  TRAIN <- read.csv("TRAIN.CSV")
} else {
  urlfile <- "https://raw.githubusercontent.com/ecpolley/
    CSMF_Data_Challenge/master/TRAIN.CSV"
  download.file(urlfile, destfile = "TRAIN.CSV")
  TRAIN <- read.csv("TRAIN.CSV")
}
dim(TRAIN)
```

```
## [1] 3000   44
```

How To Get Data

```r
# link to data on GitHub page if not available
if(file.exists("VALID.CSV")) {
  VALID <- read.csv("VALID.CSV")
} else {
  urlfile <- "https://raw.githubusercontent.com/ecpolley/
    CSMF_Data_Challenge/master/VALID.CSV"
  download.file(urlfile, destfile = "VALID.CSV")
  VALID <- read.csv("VALID.CSV")
}
dim(VALID)
```

```
## [1] 1722   43
```

Biomarkers

- Show R code

Data processing

- How are teams working with the variables?
- Did you create any new variables?
- What about missing values?

Methods

- Which prediction methods have you considered?

Loss functions

- Have you considered different loss functions other than squared error?
- What are you using to evaluate the performance of a predictor?

Beyond Analytics

- If you are working in a team with multiple individuals, how are you dividing the project?

Questions?