Project 2: Berkeley Guidance Study – Multiple Linear Regression

Due on October 5, 2015

The Berkeley Guidance Study enrolled children born in Berkeley, California, between January 1928 and June 1929, and then measured them periodically until age eighteen (Tuddenham and Snyder, 1954).

The dataset can be downloaded as `BGSall.txt` from `http://users.stat.umn.edu/~sandy/alr3ed/website/data/`, or be obtained by installing the `R` package `alr3`.

This data frame contains the following columns:

Sex	0 = males, 1 = females
WT2	Age 2 weight (kg)
HT2	Age 2 height (cm)
WT9	Age 9 weight (kg)
HT9	Age 9 height (cm)
LG9	Age 9 leg circumference (cm)
ST9	Age 9 strength (kg)
WT18	Age 18 weight (kg)
HT18	Age 18 height (cm)
LG18	Age 18 leg circumference (cm)
ST18	Age 18 strength (kg)
Soma	Somatotype, a 1 to 7 scale of body type.

1. Treat `Soma` as a continuous response variable. Use the other variables as predictors. Do multiple linear regression analysis on boys, girls, or both, respectively. Use variable transformation if necessary. For analysis on the whole dataset including `sex` as a predictor, add interaction terms between `sex` and other predictors if necessary. Report your final models.

2. Consider adding new predictors as functions of old ones. For example, check if `AVE = (WT2 + WT9 + WT18)/3`, `LIN = WT18 - WT2`, `QUAD = WT2 - 2WT9 + WT18` could help improving predictions.

3. Randomly divide the original dataset into training data (46 boys, 50 girls) and testing data (20 boys, 20 girls). The following `R` commands are recommended to use

```
set.seed(123)
training <- rep(1, 136)
training[c(sample(1:66,20),sample(67:136,20))] = 0
  # "1" indicates "training data"
BGSall.train <- BGSall[training==1,]
BGSall.test <- BGSall[training==0,]
```

Use Ridge regression, Lasso, or other shrinkage methods to check if the multiple linear regression models in part 1 can be improved. Use your training data to refit all models under comparison, and then use your testing data to compare different fitted models.

4. What is your recommended model(s) for this dataset? Does your conclusion reply on the particular partition of training vs. testing in part 3?

Notes:

[1] Students are required to work in groups on course projects and submit their reports in pdf or doc format. The group size can be 1, 2 or 3.

[2] Each group is required to submit one hard copy of the report in the class of the due day. A list of names of the group members should be put on the cover page of the report.

[3] Each report should include a nontechnical part and a technical part. The nontechnical part should be fitted into one page. The technical part should include the statement of the problem, model assumptions, formulas used, results, conclusion, suggestion and discussion if necessary.

[4] Any statistical software and other tools can be used. If R program is used, the attachment of R code is recommended.