

STAT406 - Fall 2017 - Lecture # 3

Matías Salibián Barrera

14 Sep 2017

1 IN-CLASS ACTIVITY

- (1) Consider the `fallacy.dat` data set. This is a synthetic data set that has been generated as follows:

```
n <- 150
p <- 110
set.seed(39085)
y <- rnorm(n, mean = mu, sd=1)
x <- matrix( rnorm(n*p), n, p)
dat <- data.frame(list(y=y, x=x))
```

- (2) What is the distribution of X_1, X_2, \dots , and X_{110} ? What is the true regression function $E[Y|X_1, \dots, X_{110}]$?
- (3) What do you think should be the best linear regression predictor for future Y 's?
- (4) Use a forward or backward stepwise method to select a subset of predictors that produce a good linear regression fit. Note that you can use the function `stepAIC` in the R package `MASS` to perform stepwise variable selection.
- (5) Use 10 runs of a simple 5-fold cross-validation method to estimate the mean squared prediction error of:

- (a) the model found in (4), and
 - (b) the predictor in (3).
- (6) Find (using the inequality in the lecture notes) a lower bound for the true mean squared prediction error

$$E_{\text{data},(Y,\mathbf{X})} \left[(Y - \hat{f}(\mathbf{X}))^2 \right],$$

of *any* predictor \hat{f} for this model, where $\mathbf{X} = (X_1, \dots, X_{110})^\top$. What does this lower bound tell you about the results in (5)?

- (7) What are we doing wrong in (5)?
- (8) Re-do (5) correctly. Are these new results compatible with the bound found in (6)?
- (9) Consider the Air Pollution data set used in class. Use 10 runs of 5-fold CV to compare the mean squared prediction errors of the linear model chosen via forward stepwise and that of the full model.