Inference after model selection in high dimensional linear regression

Lecture 1

Richard Lockhart, Simon Fraser University

University of Cambridge: Mini-course, Lent term, 2017

January 23, 2017
Outline

- Illustrative example.
Example for Context

- riboflavin data set from hdi package in R.
- See Bühlmann et al. [2014]
- \( n = 71 \) values of log production of riboflavin, \( Y \).
- \( p = 4088 \) covariates: expression levels of 4088 genes, \( X \).
- Linear model as usual:
  \[
  Y = \beta_0 + X\beta + \epsilon
  \]
- Homoscedastic mean 0 errors with variance \( \sigma^2 \).
- Model is not identified without restrictions on \( \beta \).
All 4008 correlations with $Y$
Scatterplots, $Y$ and top 6 correlations
Histogram of all 4088 correlations