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.

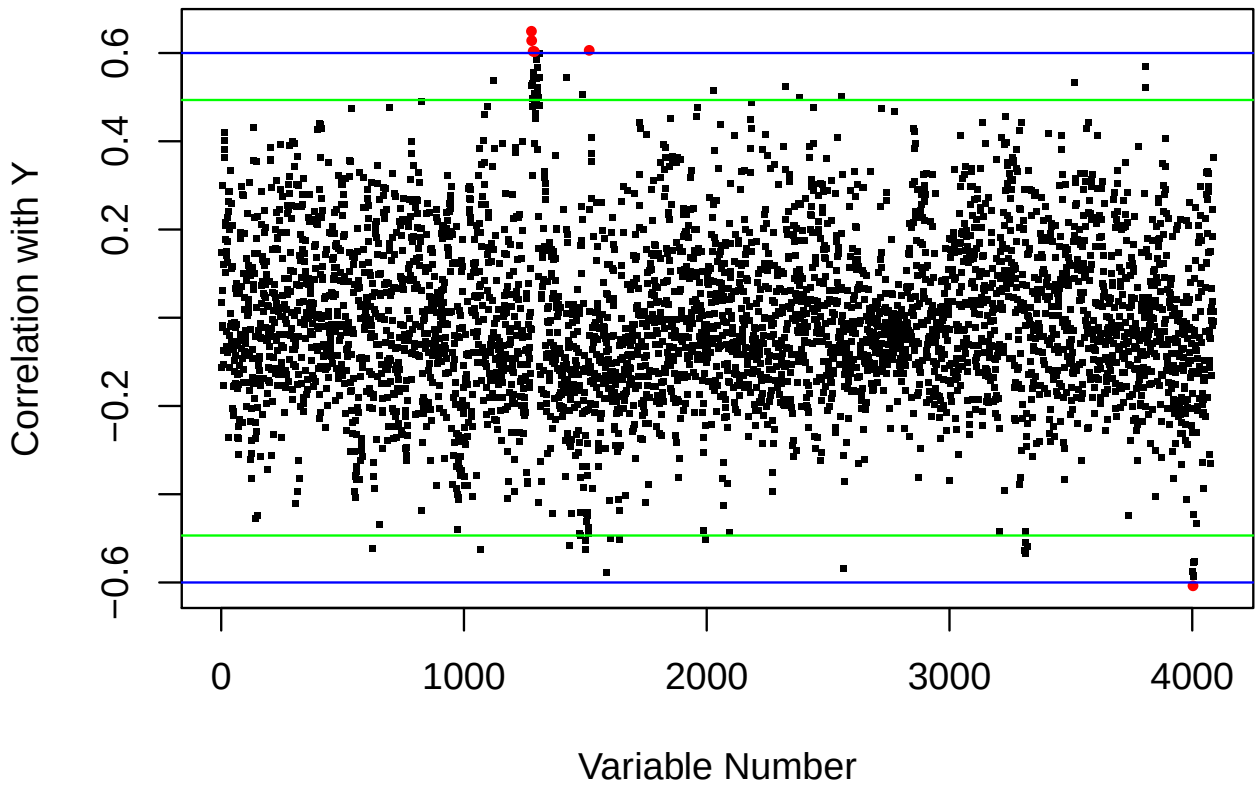
Example for Context

- `riboflavin` data set from `hdi` package in `R`.
- See Bühlmann et al. [2014]
- $n = 71$ values of log production of riboflavin, $\mathbf{Y}$.
- $p = 4088$ covariates: expression levels of 4088 genes, $\mathbf{X}$.
- Linear model as usual:

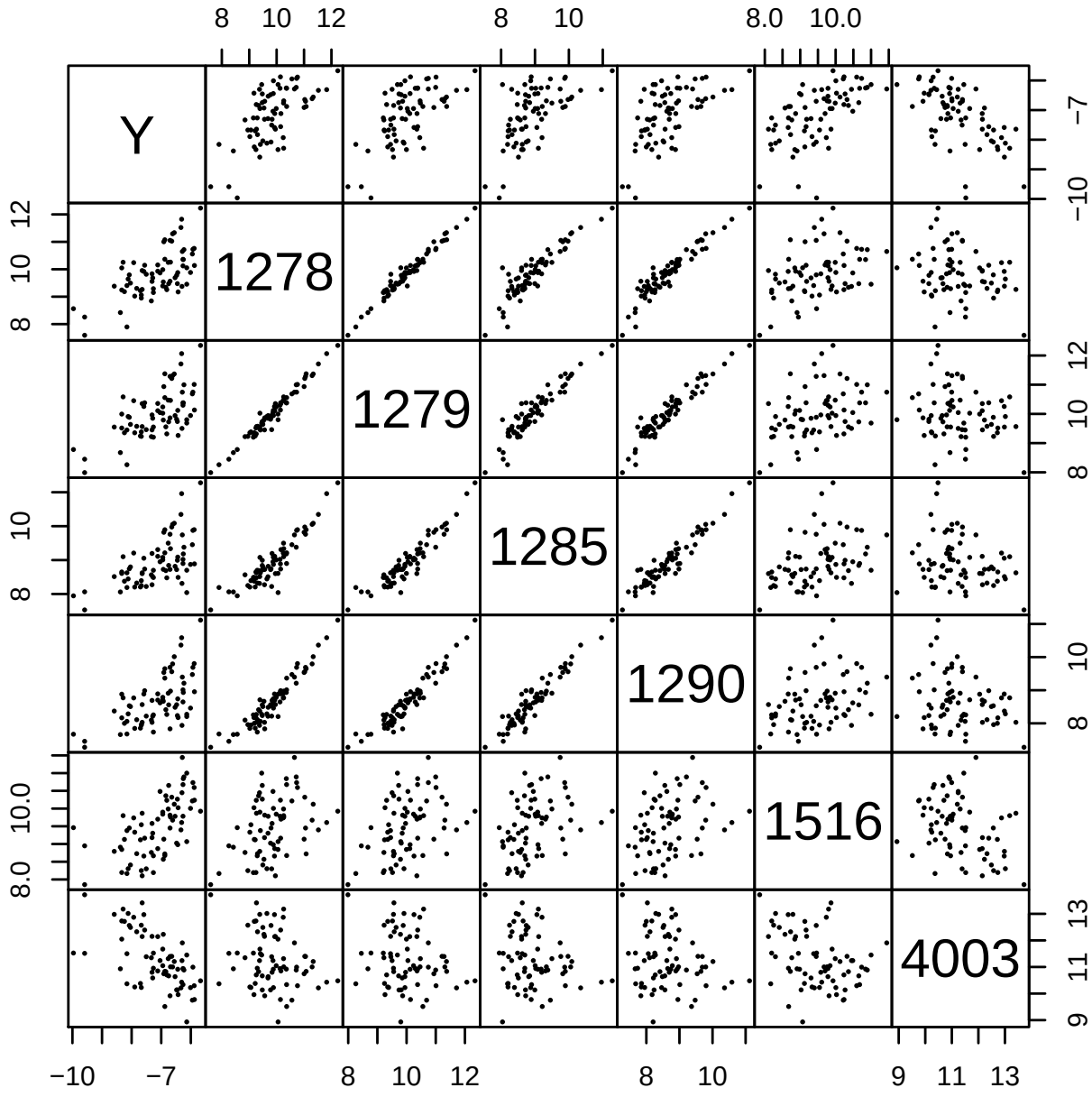
$$\mathbf{Y} = \beta_0 + \mathbf{X}\beta + \epsilon$$

- Homoscedastic mean 0 errors with variance σ^2.
- Model is not identified without restrictions on β.

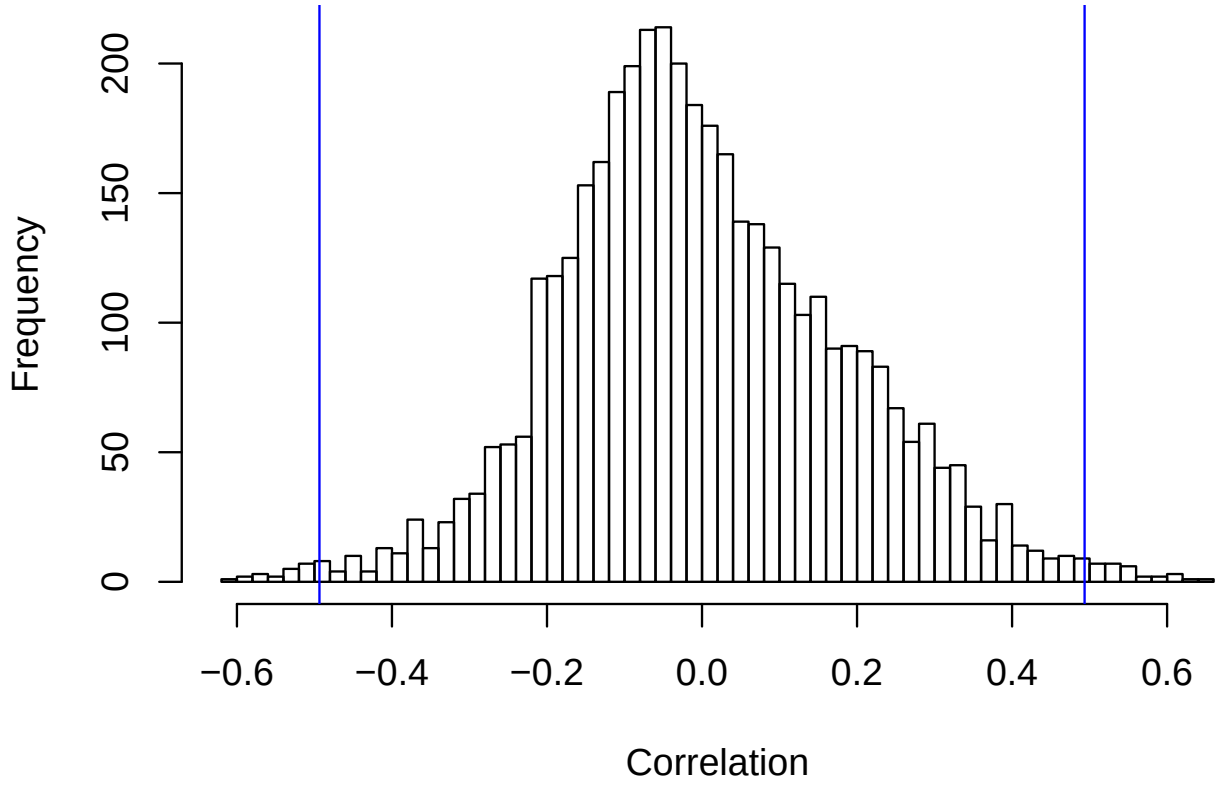
All 4008 correlations with Y



Scatterplots, Y and top 6 correlations



Histogram of all 4088 correlations



References

Peter Bühlmann, Markus Kalisch, and Lukas Meier. High-dimensional statistics with a view toward applications in biology. *Annual Review of Statistics and Its Application*, 1(1): 255–278, 2014. doi: 10.1146/annurev-statistics-022513-115545. URL /brokenurl#http://dx.doi.org/10.1146/annurev-statistics-022513-115545.