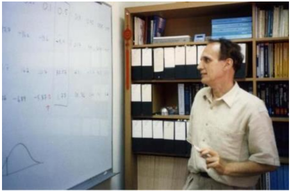
Prediction, Estimation, and Attribution

Statistical Learning
CLAMSES - University of Milano-Bicocca

Aldo Solari



Bradley Efron working in his classic office, circa 1996.

References

- International Prize in Statistics 2019
- Efron, B. (2020). Prediction, Estimation, and Attribution. Journal of the American Statistical Association, 115(530), 636-655. With Discussion and Rejoinder.
- Slides
- Recorded presentation for the 62nd ISI World Statistics Congress in Kuala Lumpur [46 mins]

Outline

1. Introduction
2. Surface Plus Noise Models
3. The Pure Prediction Algorithms
4. A Microarray Prediction Problem
5. Advantages and Disadvantages of Prediction
6. The Training/Test Set Paradigm
7. Smoothness
8. A Comparison Checklist
9. TraditionalMethods in the Wide Data Era
10. Two Hopeful Trends

Regression
Gauss (1809), Galton (1877)

What are the three important statistical tasks in regression?

- *Prediction: the prediction of new cases*
  e.g. random forests, boosting, support vector machines, neural nets, deep learning
- *Estimation: the estimation of regression surfaces*
  e.g. OLS, logistic regression, GLM: MLE
- *Attribution: the assignment of significance to individual predictors*
  e.g. ANOVA, lasso, Neyman Pearson

How do the pure prediction algorithms relate to traditional regression methods?

That is the central question pursued in what follows.

2. Surface Plus Noise Models

We will assume that the data $\mathbf{d}$ available to the statistician has this structure:

$$\mathbf{d} = \{(x_i, y_i), i = 1, \ldots, n\}$$

- x_i is a p-dimensional vector of predictors taking its value in a known space $\mathcal{X}$ contained in $\mathbb{R}^p$;
- y_i is a real valued response;
- the n pairs are assumed to be independent of each other.

More concisely we can write

$$\mathbf{d} = \{\mathbf{x}, \mathbf{y}\}$$

where $\mathbf{x}$ is the $n \times p$ matrix having x_i^t as the ith row, and $\mathbf{y} = (y_1, \ldots, y_n)^t$.

- The regression model is

$$y_i = s(x_i, \beta) + \epsilon_i \quad i = 1, \ldots, n \tag{1}$$

$\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ where $s(x, \beta)$ is some functional form that, for any fixed value of the parameter vector β, gives expectation $\mu = s(x, \beta)$ as a function of $x \in \mathcal{X}$;

- The *regression surface* is

$$\mathcal{S}_\beta = \{\mu = s(x, \beta), x \in \mathcal{X}\}$$

Most traditional regression methods depend on some sort of surface plus noise formulation;

- The surface describes the scientific truths we wish to learn, but we can only observe points on the surface obscured by noise;
- The statistician's traditional estimation task is to learn as much as possible about the surface from the data **d**.

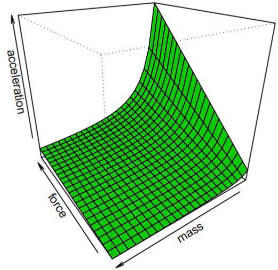
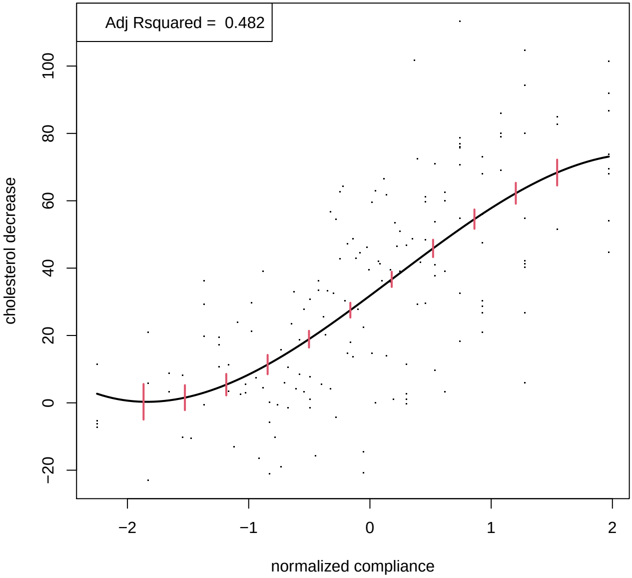


Figure 2. On left, a surface depicting Newton's second law of motion, acceleration = force/mass; on right, a noisy version.

Cholesterol data

- Cholestyramine, a proposed cholesterol lowering drug, was administered to 164 men for an average of seven years each.
- The response variable is a man's decrease in cholesterol level over the course of the experiment.
- The single predictor is compliance, the fraction of intended dose actually taken (standardized)
- https://hastie.su.domains/CASI_files/DATA/cholesterol.html



Adj Rsquared = 0.482
cholesterol decrease
−20
0
20
40
60
80
100
−2
−1
0
1
2
normalized compliance