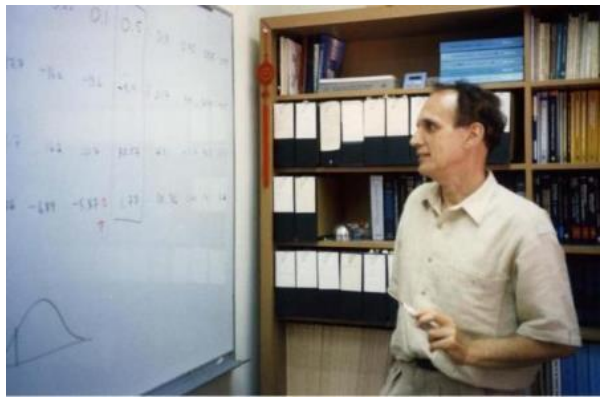
Prediction, Estimation, and Attribution

Statistical Learning

CLAMSES - University of Milano-Bicocca

Aldo Solari



Bradley Efron working in his classic office, circa 1996.

References

This material reproduces the following

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

Table of Contents

Regression

Gauss (1809), Galton (1877)

- *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. Fisher's ANOVA, Neyman-Pearson

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

That is the central question pursued in what follows.

Table of Contents

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

$$\mathcal{D} = \{(x_i, y_i), i = 1, \ldots, n\}$$

- 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

$$\mathcal{D} = \{X, y\}$$

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

Regression surface

- The regression model is

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

  $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$ 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

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

  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 $\mathcal{D}$.

The left panel of the Figure shows the surface representation of Newton's second law of motion,

acceleration = force / mass

The right panel shows a picture of what experimental data might have looked like.

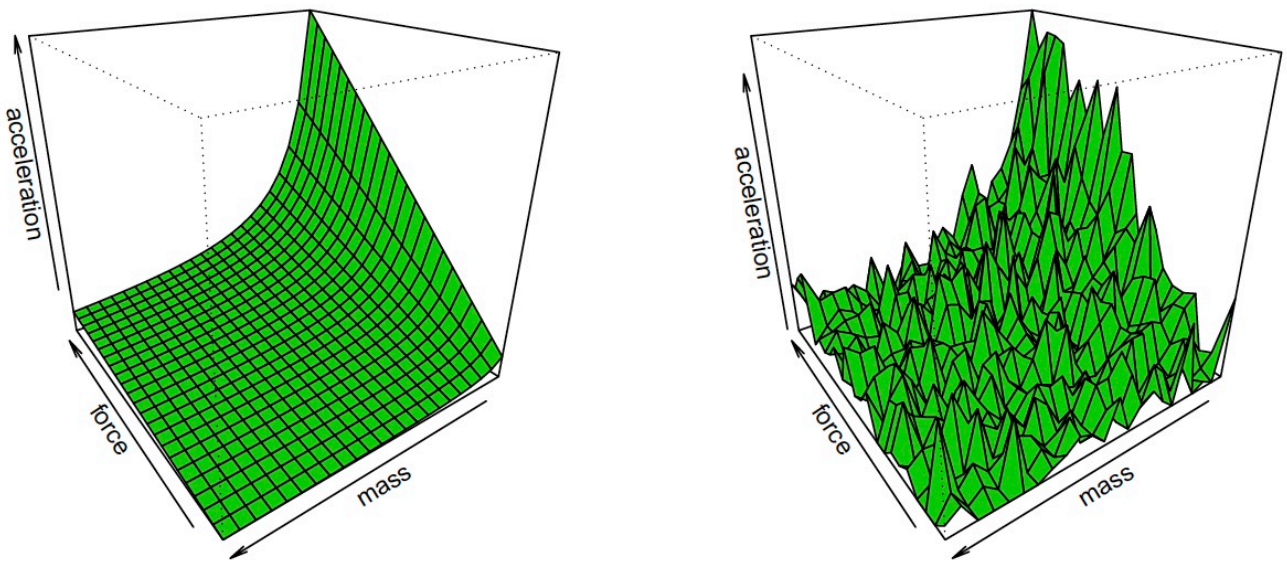


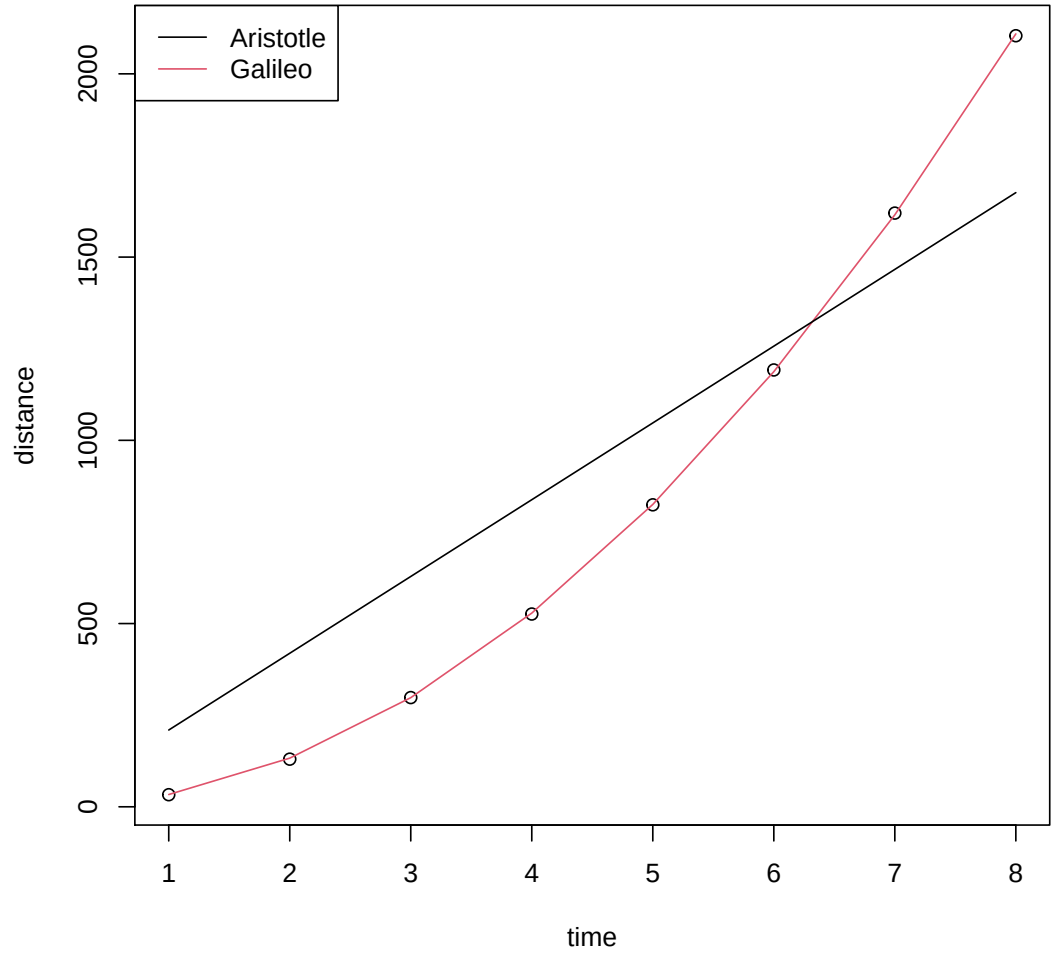
Figure 2. On left, a surface depicting Newton's second law of motion, acceleration = force/mass; on right, a noisy version.

Galileo's inclined plane experiment (1604)

- If a ball rolls down a ramp, what is the relationship between time (x) and distance (y)?
- Aristotle: Constant velocity (zero acceleration): distance $\propto$ time
- Galileo : Increasing velocity (constant acceleration): distance $\propto$ time^2
- Experimental data:

time	1	2	3	4	5	6	7	8
distance	33	130	298	526	824	1192	1620	2104

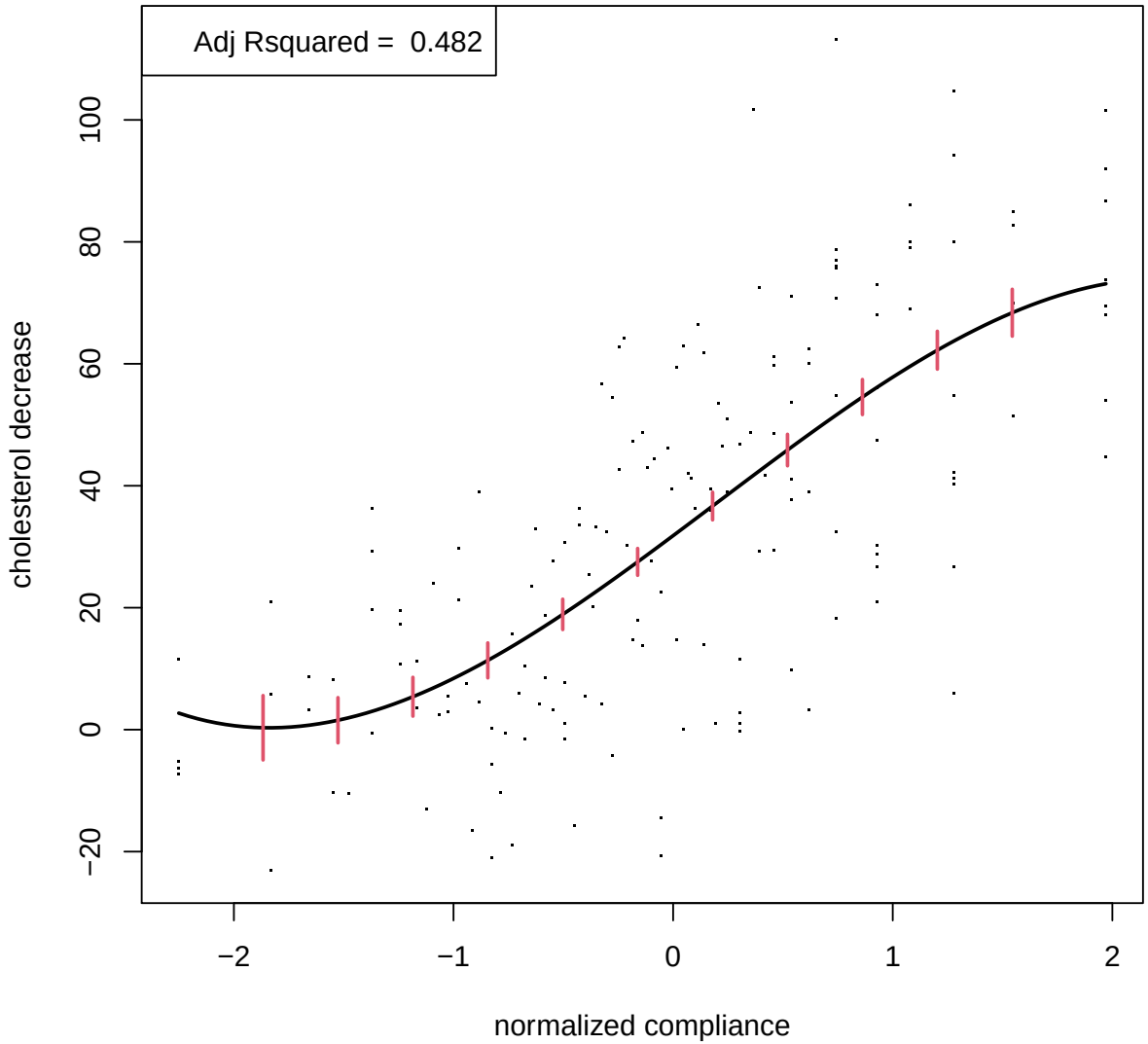
MacDougal, D. W. (2012). Galileo's Great Discovery: How Things Fall. In Newton's Gravity (pp. 17-36). Springer



https://github.com/aldosolari/SL/blob/master/docs/RCODE/EfronPEA.R

Cholesterol data

- Cholestyramine, a proposed cholesterol lowering drug, was administered to 164 male doctors for an average of seven years each (Efron and Feldman, 1991)
- The response variable (y_i) is a man's decrease in cholesterol level over the course of the experiment.
- The single predictor is compliance (x_i), the fraction of intended dose actually taken. Compliance, the proportion of the intended dose actually taken, ranged from 0% to 100%, −2.25 to 1.97 on the normalized scale. It was hoped to see larger cholesterol decreases for the better compliers.
- https://hastie.su.domains/CASI_files/DATA/cholesterol.html


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

- A normal regression model was fit, with

$$s(x_i, \beta) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3$$

  in other words, a cubic regression model.
- The black curve is the estimated surface

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

  fit by maximum likelihood or, equivalently, by ordinary least squares (OLS).
- The vertical bars indicate one standard error for the estimated values $s(x, \hat{\beta})$, at 11 choices of x, showing how inaccurate $\hat{\mathcal{S}}$ might be as an estimate of the true $\mathcal{S}$
- Only $\hat{\beta}_0$ and $\hat{\beta}_1$ were significantly nonzero. The adjusted R^2 was 0.482, a traditional measure of the model's predictive power.

Table of Contents

- Random Forests, Boosting, Deep Learning, etc.
- Data
$$\mathcal{D} = \{(x_i, y_i), i = 1, \ldots, n\}$$
- Prediction rule $f(x, \mathcal{D})$
- New $(x, ?)$ gives $\hat{y} = f(x, \mathcal{D})$
- Strategy: Go directly for high predictive accuracy; forget (mostly) about surface + noise

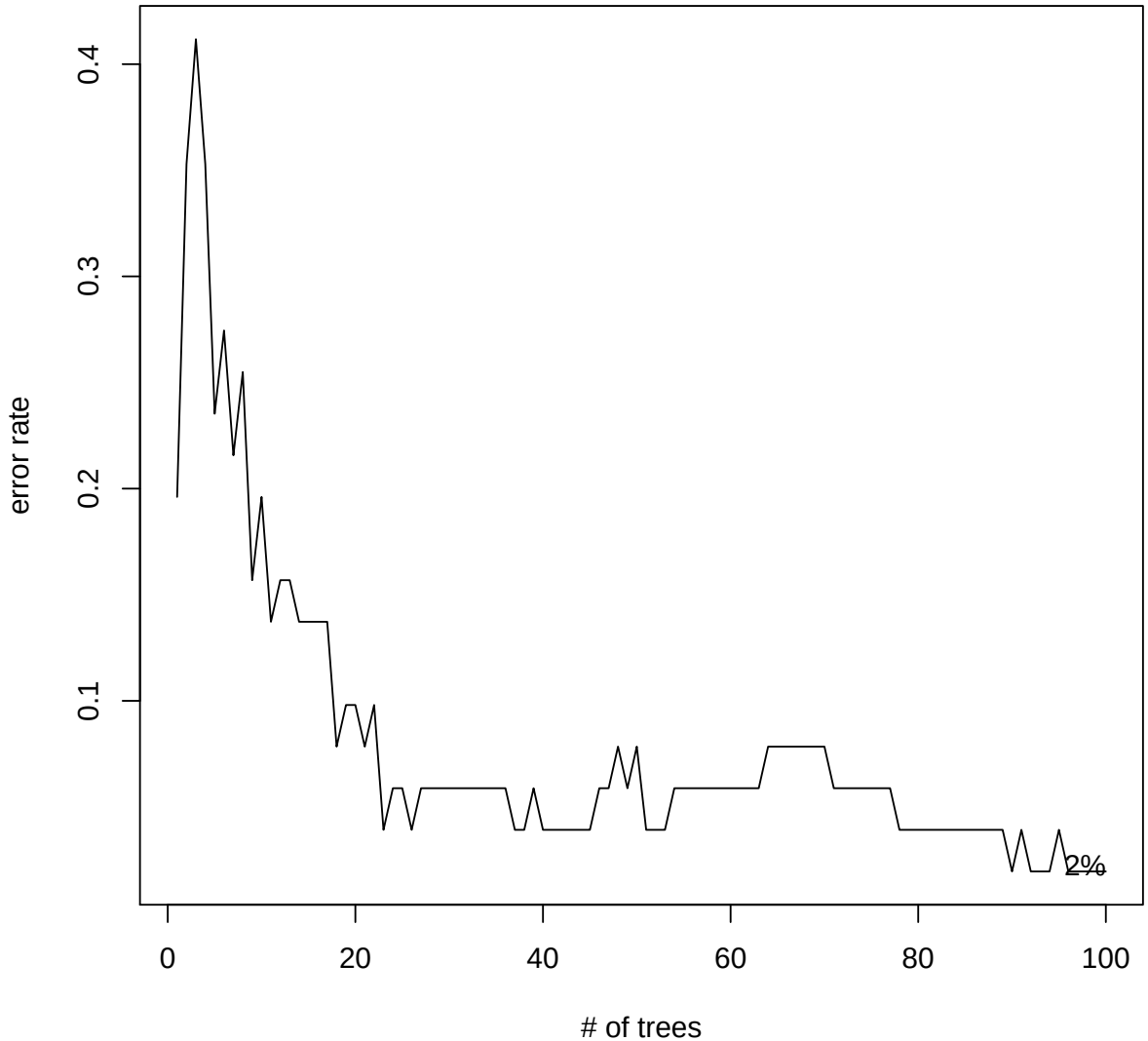
Table of Contents

The Prostate Cancer Microarray Study

- https://hastie.su.domains/CASI_files/DATA/prostate.html
- $n = 102$ men: 52 prostate cancer, 50 normal controls
- For each man measure activity of $p = 6033$ genes
- Data set D is 102×6033 matrix ("wide")
- Wanted: Prediction rule $f(x, \mathcal{D})$ that inputs new 6033-vector x and outputs $\hat{y}$ correctly predicting cancer/normal

Random forest

- Randomly divide the 102 subjects into:
  - training set of 51 subjects (26 + 25)
  - test set of 51 subjects (26 + 25)
- Run R program `randomForest` on the training set
- Use its rule $f(x_i, D)$ on the test set and see how many errors it makes


error rate
0.1
0.2
0.3
0.4
0
20
40
60
80
100
# of trees
2%

Boosting

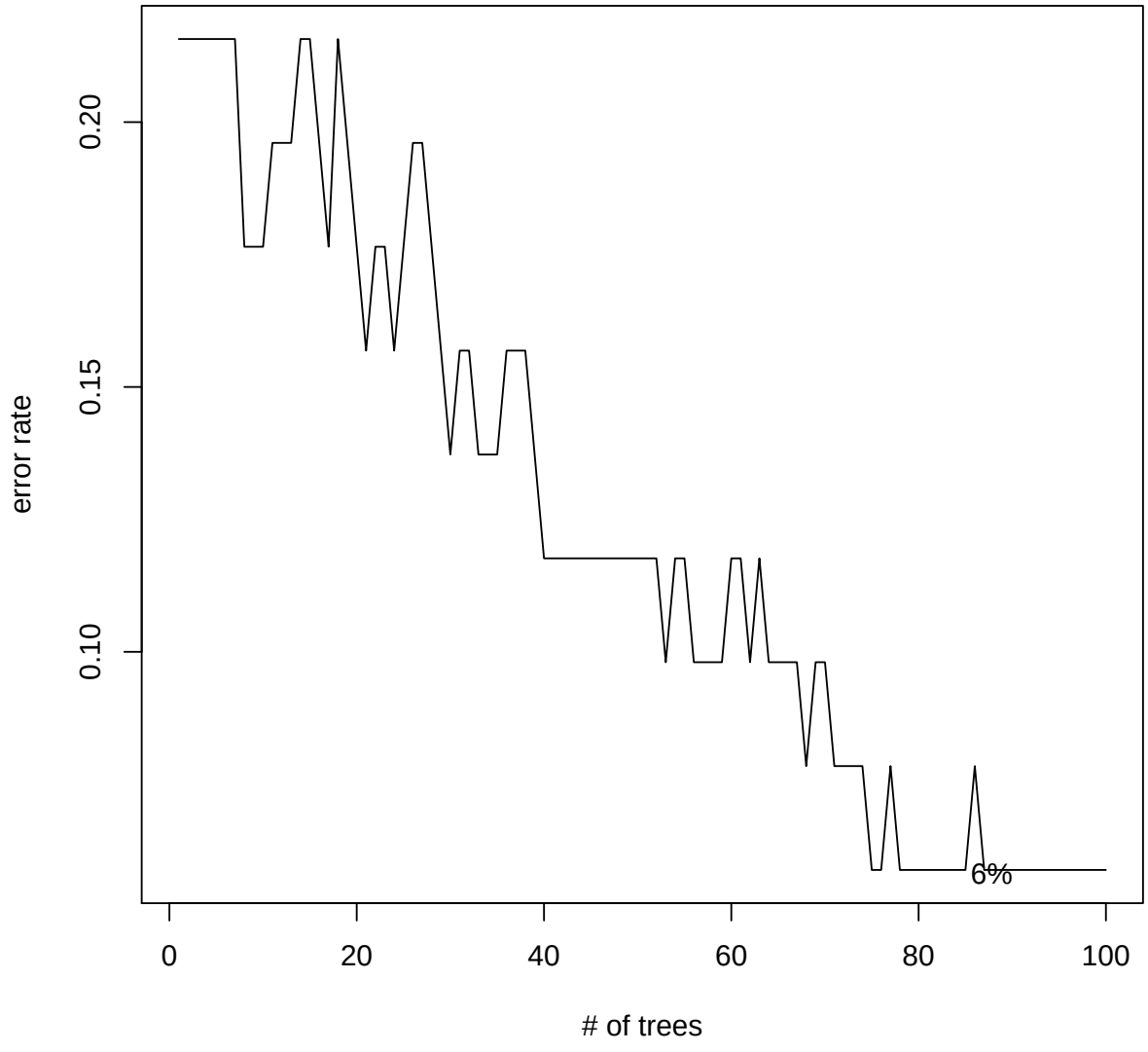
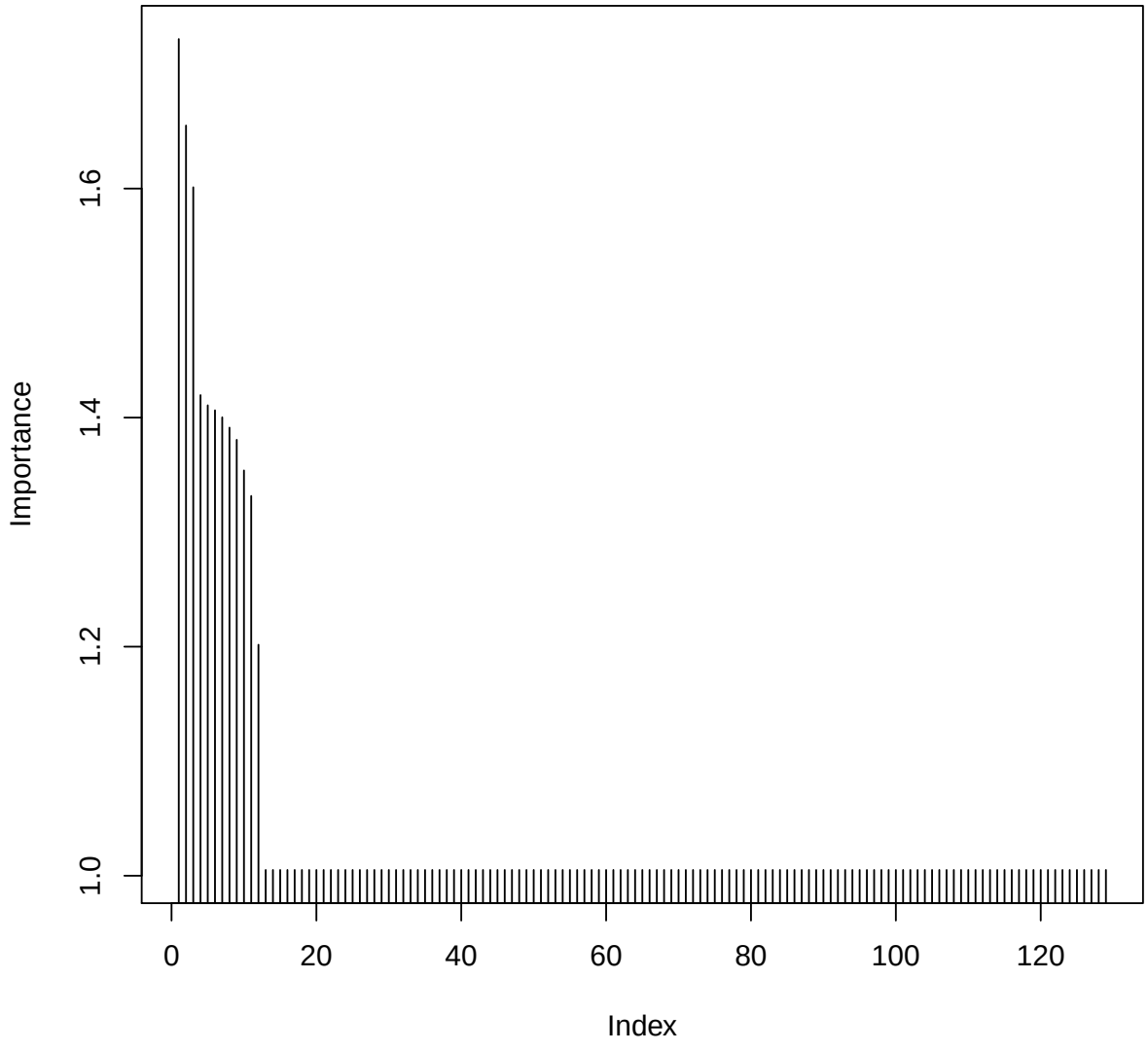


Table of Contents

Variable importance

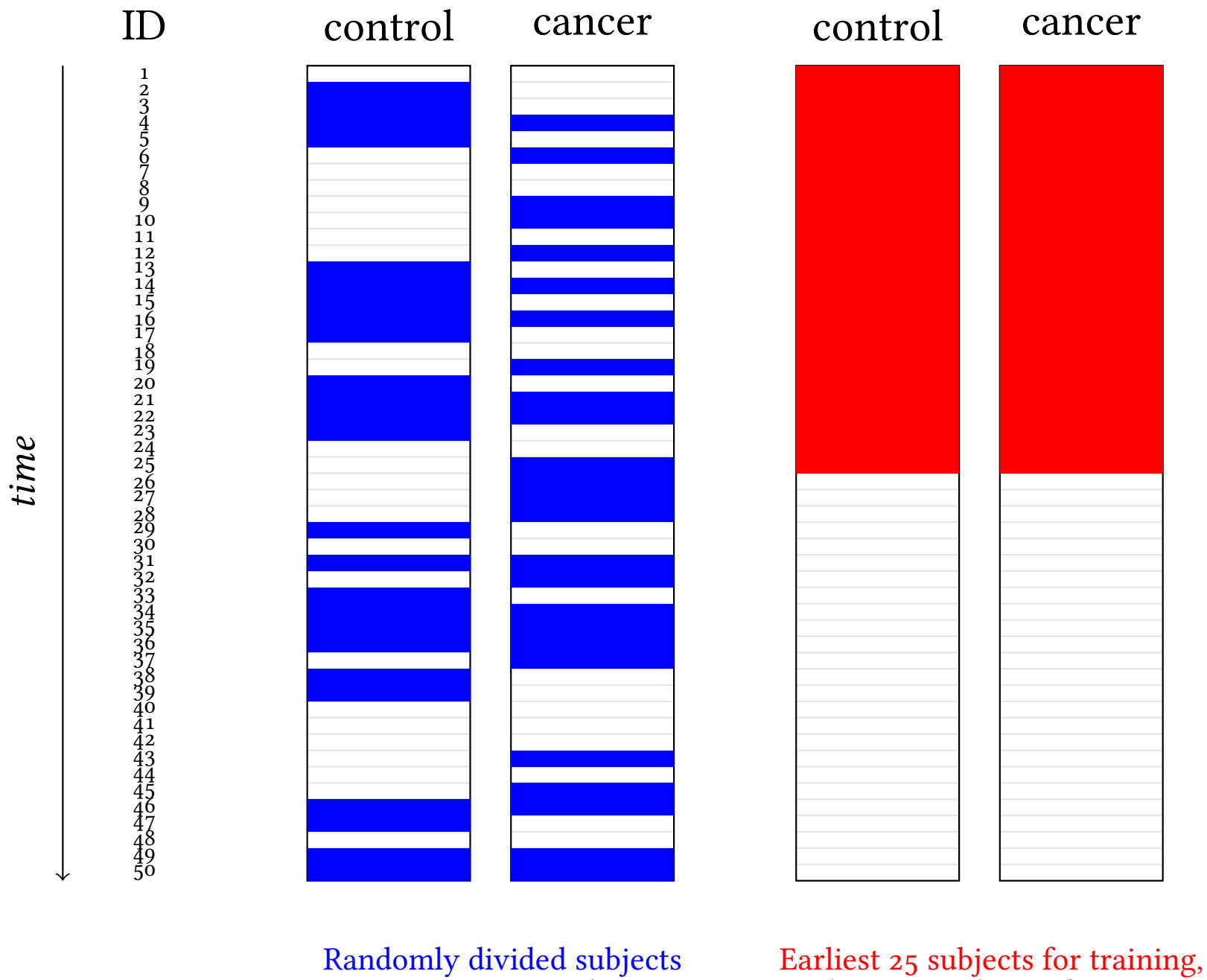


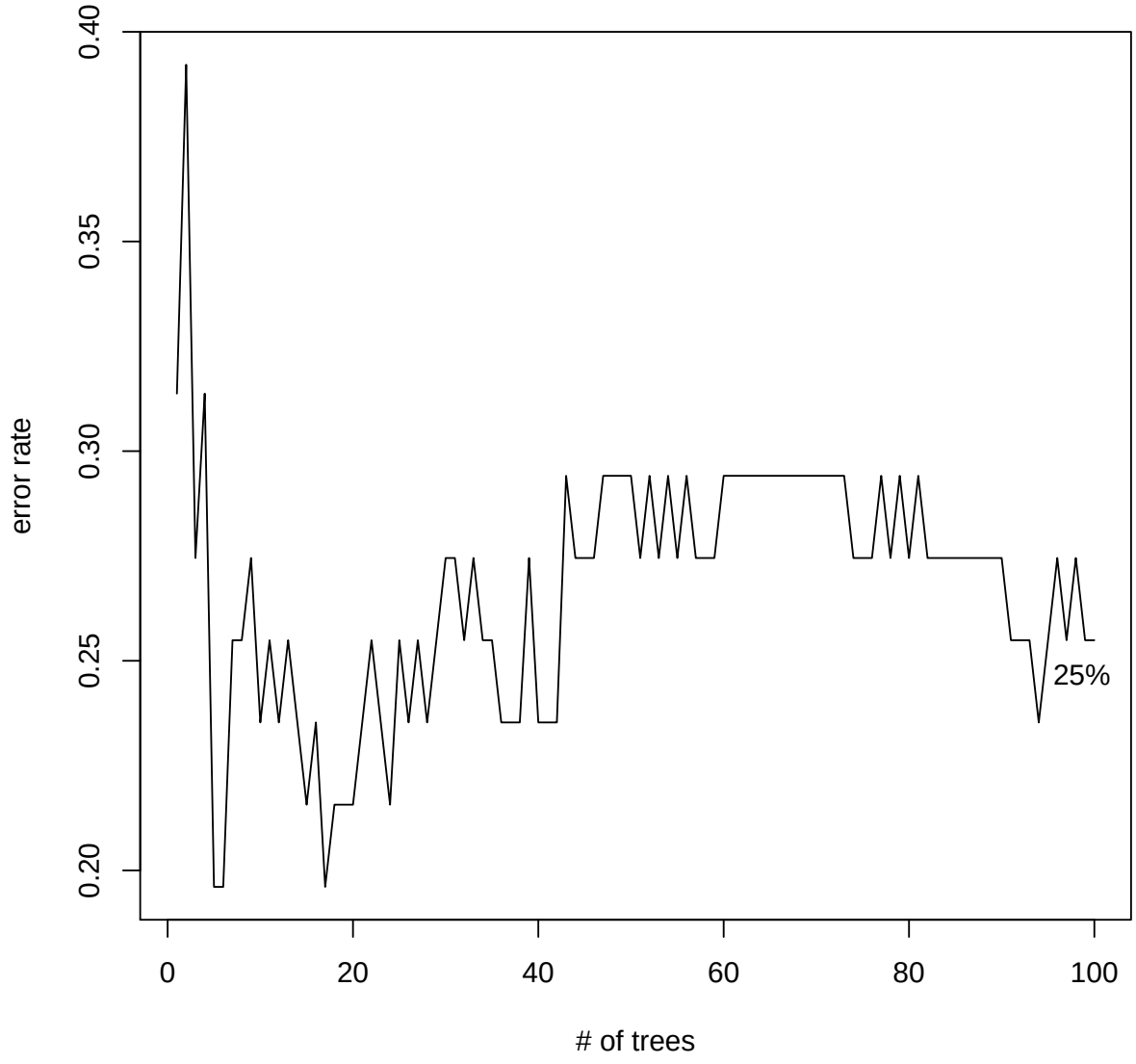
- Importance measure is computed for each of the p predictor variables.
- Of the $p = 6033$ genes, 129 had positive scores, these being the genes that ever were chosen as splitting variables.
- Can we use the importance scores for attribution?
- The answer seems to be no. Removing the most important 100 had similarly minor effects on the number of test set prediction errors
- Evidently there are a great many genes weakly correlated with prostate cancer, which can be combined in different combinations to give near-perfect predictions.

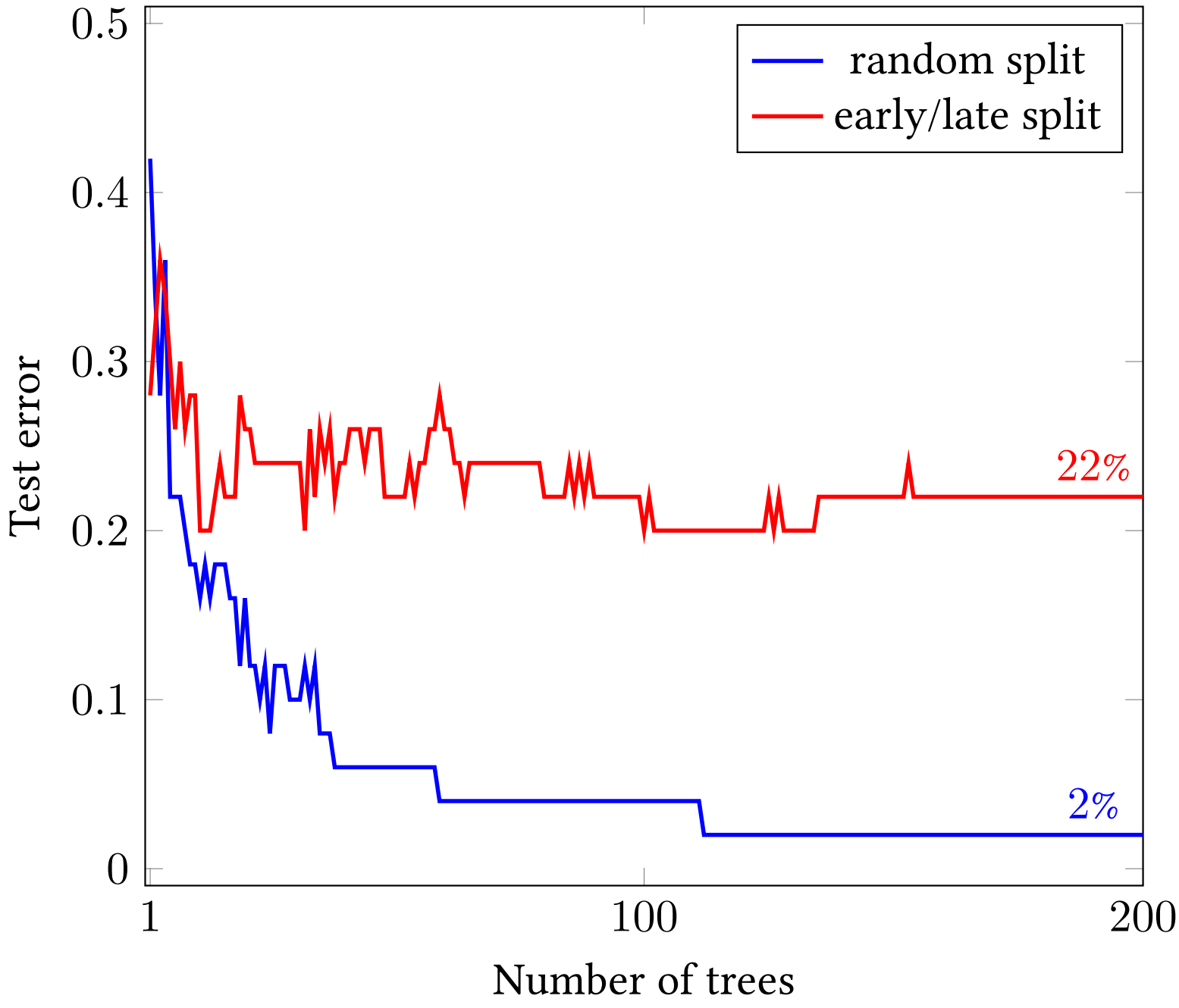
Table of Contents

Were the Test Sets Really a Good Test?

- Prediction can be highly context-dependent and fragile
- Before Randomly divided subjects into training and test
- Next:
  - 51 earliest subjects for training (25 control + 26 cancer with lowest ID numbers)
  - 51 latest subjects for test
- Study subjects might have been collected in the order listed, with some small methodological differences creeping in as time progressed (concept drift)


ID
control
cancer
control
cancer
time
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
Randomly divided subjects into training and test
Earliest 25 subjects for training, latest 25 subjects for test


error rate
0.20
0.25
0.30
0.35
0.40
0
20
40
60
80
100
# of trees
25%


0.5
0.4
0.3
0.2
0.1
0
Test error
1
100
200
Number of trees
random split
early/late split
22%
2%

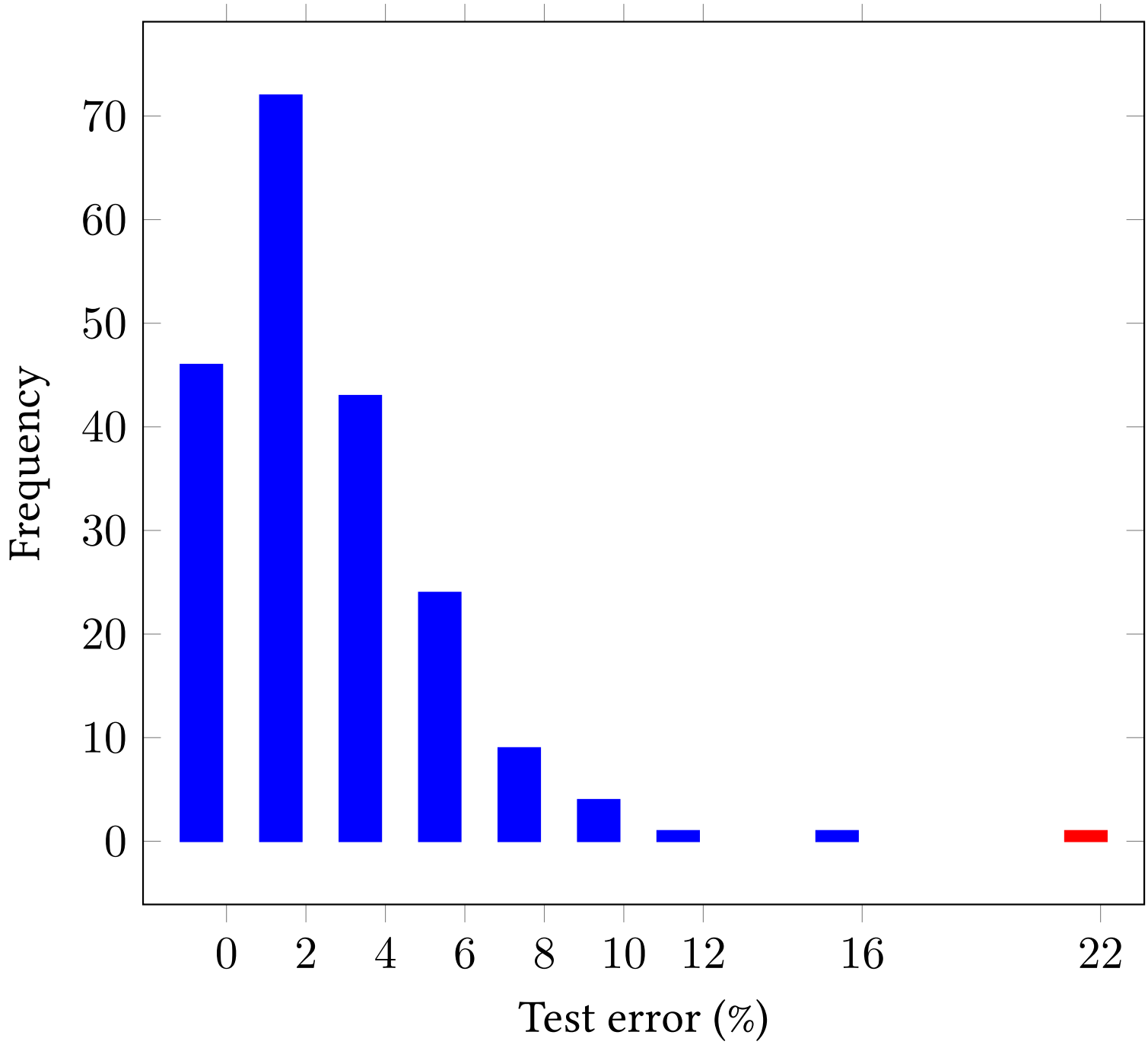
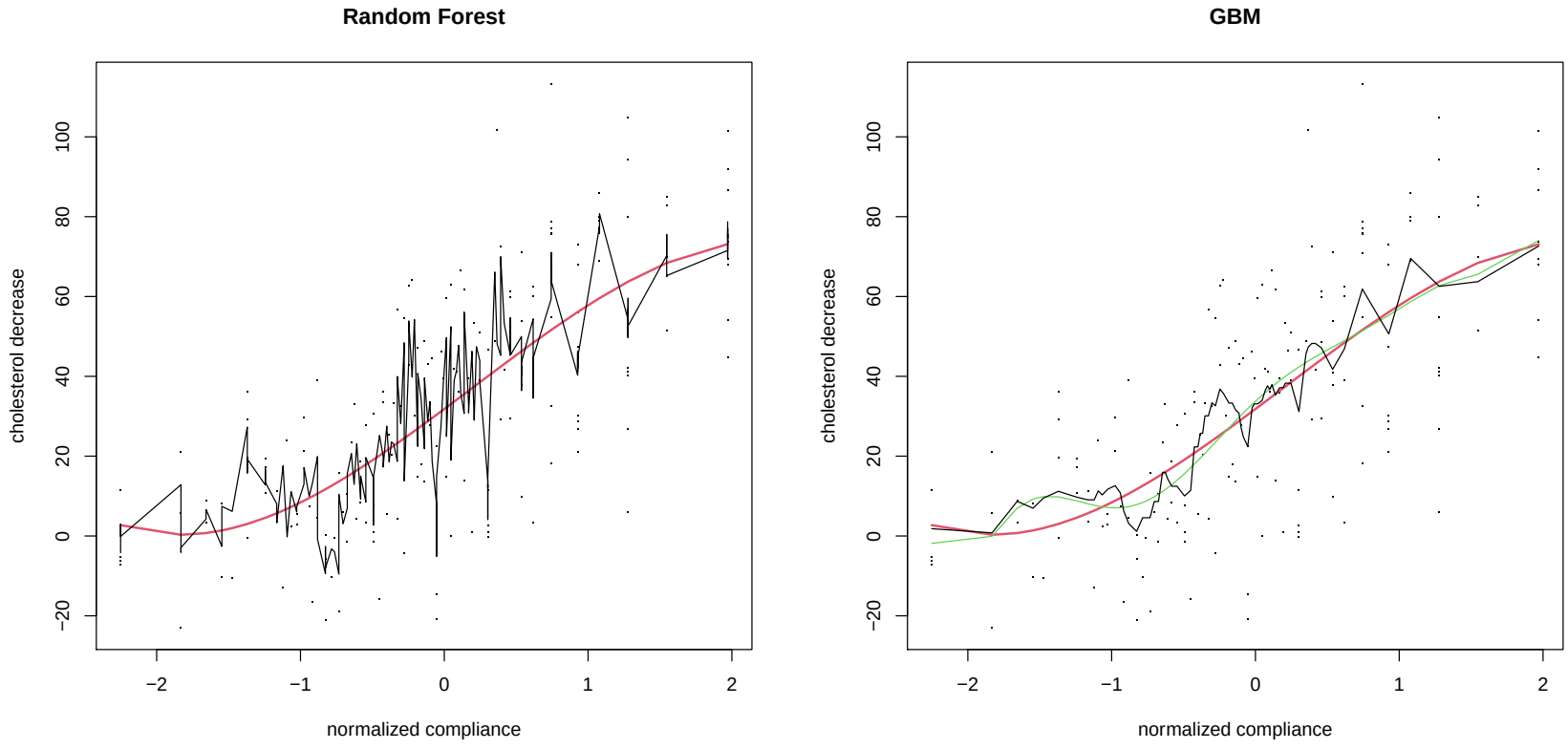

Frequency
70
60
50
40
30
20
10
0
0
2
4
6
8
10
12
16
22
Test error (%)

Table of Contents

- The parametric models of traditional statistical methodology enforce the smooth-world paradigm
- Looking back at the Cholesterol data, we might not agree with the exact shape of the cholostyramine cubic regression curve but the smoothness of the response seems unarguable
- The choice of cubic was made on the basis of a Cp comparison of polynomial regressions degrees 1 through 8, with cubic best.
- Smoothness of response is not built into the pure prediction algorithms.
- Random forest and algorithm `gbm` take X to be the 164×8 matrix `poly(c,8)` - an 8th degree polynomial basis



`randomForest` and `gbm` fits to the Cholesterol data. Heavy red curve is cubic OLS; dashed green curve in right panel is 8th degree OLS fit.

Table of Contents

	Traditional regressions methods	Pure prediction algorithms
1.	Surface plus noise models (continuous, smooth)	Direct prediction (possibly discrete, jagged)
2.	Scientific truth (long-term)	Empirical prediction accuracy (possibly short-term)
3.	Parametric modeling (causality)	Nonparametric (black box)
4.	Parsimonious modeling (researchers choose covariates)	Anti-parsimony (algorithm chooses predictors)
5.	X $n \times p$ with $p \ll n$ (homogeneous data)	$p \gg n$, both possibly enormous (mixed data)
6.	Theory of optimal inference (mle, Neyman–Pearson)	Training/test paradigm (Common Task Framework)

Table of Contents

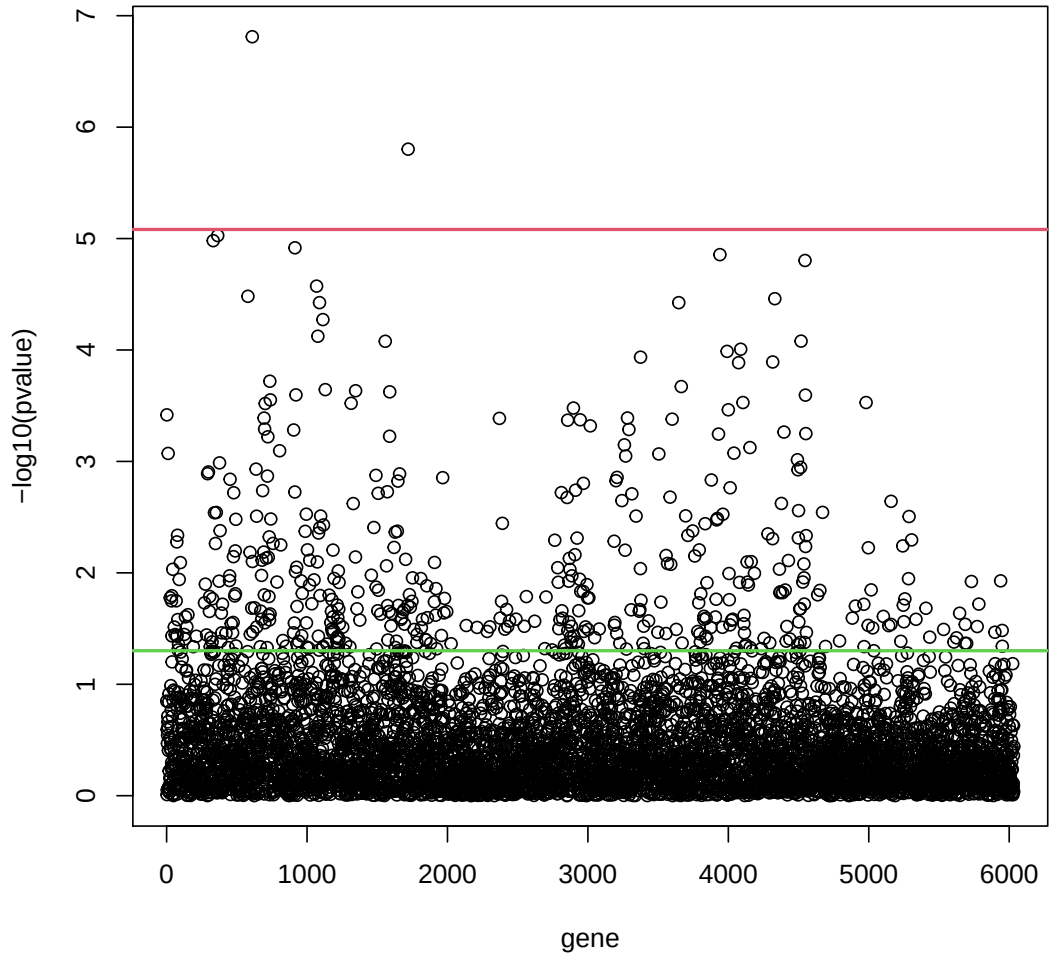
Estimation and Attribution in the Wide-Data Era

- Large p (the number of features) affects Estimation
  - MLE can be badly biased for individual parameters
  - "surface" if, say, $p = 6033$?
- Attribution still of interest. Compute p-value p_i for the null hypothesis H_i: no difference in gene expression between cancer and control at the ith gene
- The Bonferroni threshold for 0.05 significance is

$$p_i \leq 0.05/6033$$

$$\begin{aligned} \Pr(\text{at least one Type I error}) &= \Pr\left(\bigcup_{i \in I_0} \{p_i \leq \alpha/p\}\right) \\ &\leq \sum_{i \in I_0} \mathrm{P}(p_i \leq \alpha/p) \leq |I_0| \frac{\alpha}{p} \leq \alpha \end{aligned}$$

- Instead of performing a traditional attribution analysis with $p = 6033$ predictors, a microarray analysis performs 6033 analyses with $p = 1$


−log10(pvalue)
0
1
2
3
4
5
6
7
0
1000
2000
3000
4000
5000
6000
gene

- Sparsity offers another approach to wide-data estimation and attribution: we assume that most of the p predictor variables have no effect and concentrate effort on finding the few important ones.
- The lasso provides a key methodology. Estimate β, the p-vector of regression coefficients, by minimizing

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^t\beta) + \lambda\|\beta\|_1$$

  where $\|\beta\|_1 = \sum_{j=1}^{p}|\beta_j|$
- Here λ is a fixed tuning parameter: $\lambda = 0$ corresponds to the OLS solution for β (if $p \leq n$) while $\lambda = \infty$ makes $\hat{\beta} = 0$. For large values of λ only a few of the coordinates $\hat{\beta}_j$ will be nonzero.
- The lasso produced biased estimates of β, with the coordinate values $\hat{\beta}_j$ shrunk toward zero.

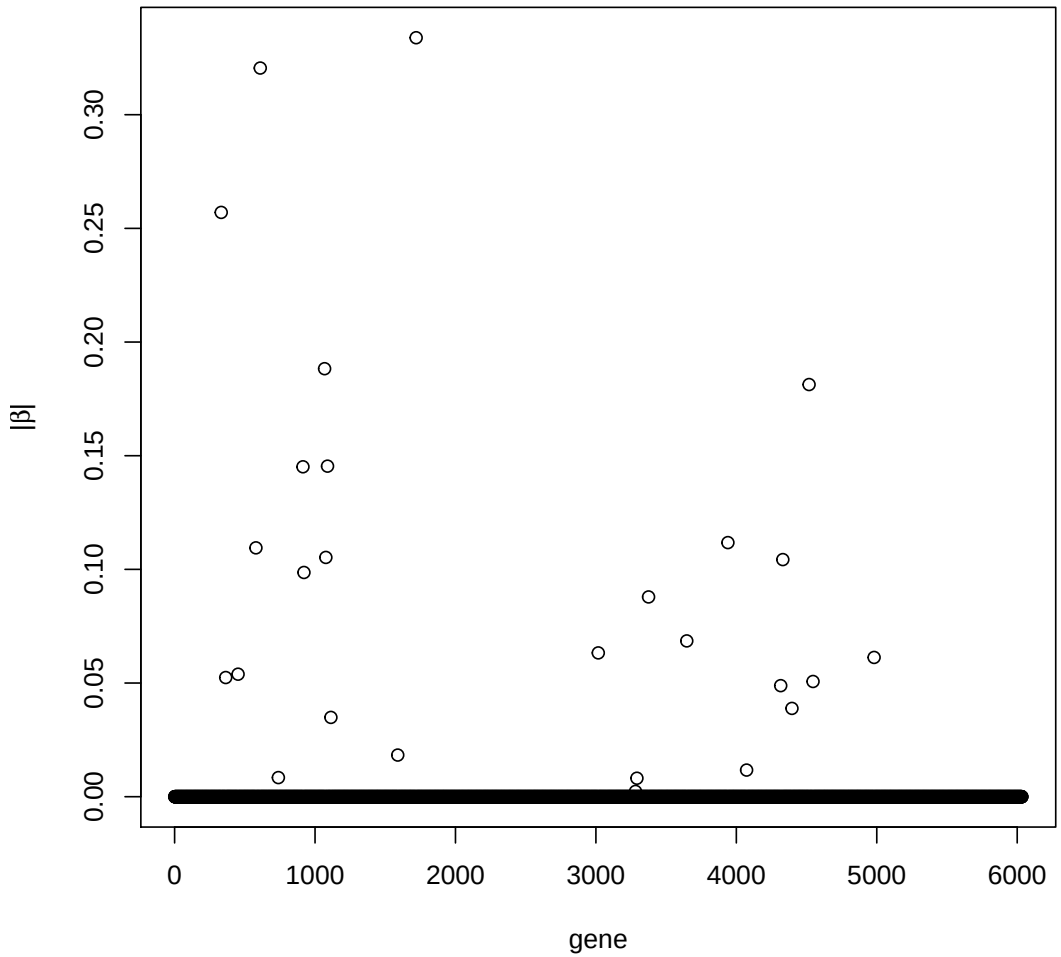

|β|
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0
1000
2000
3000
4000
5000
6000
gene

Table of Contents

- Making prediction algorithms better for scientific use
  - smoother
  - more interpretable
- Making traditional estimation/attribution methods better for large-scale (n, p) problems
  - more flexible
  - better scaled
- We do have optimality theory for estimation (MLE) and attribution (Neyman-Pearson), but we do not have an optimality theory for prediction.