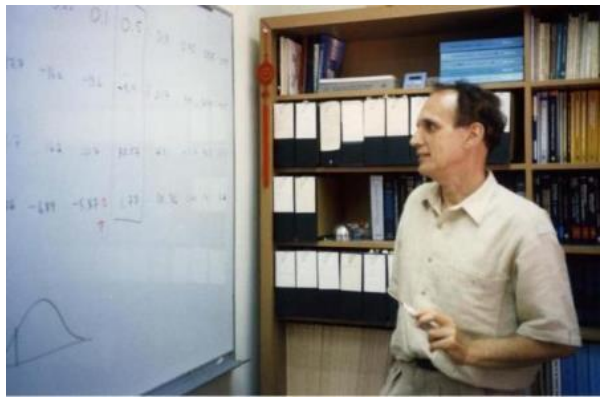
Prediction, Estimation, and Attribution

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 $\mathbf{d}$.

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

acceleration = force / mass

It is pleasing to imagine the second law falling full-born out of Newton's head, but he was a master of experimentation. The right panel shows a (fanciful) picture of what experimental data might have looked like.

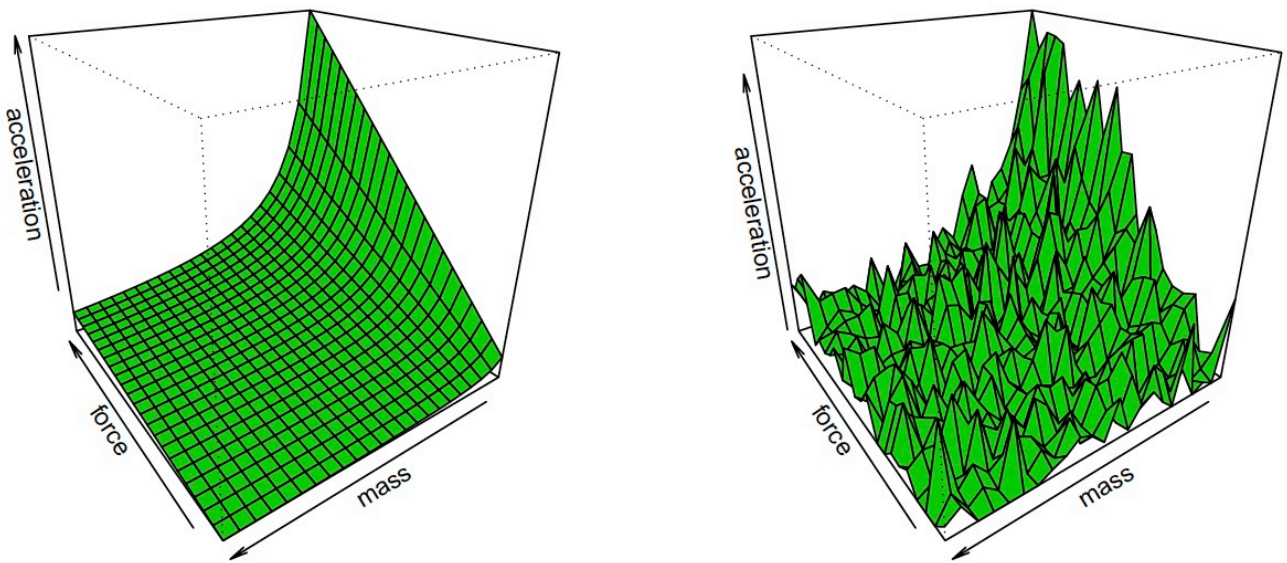
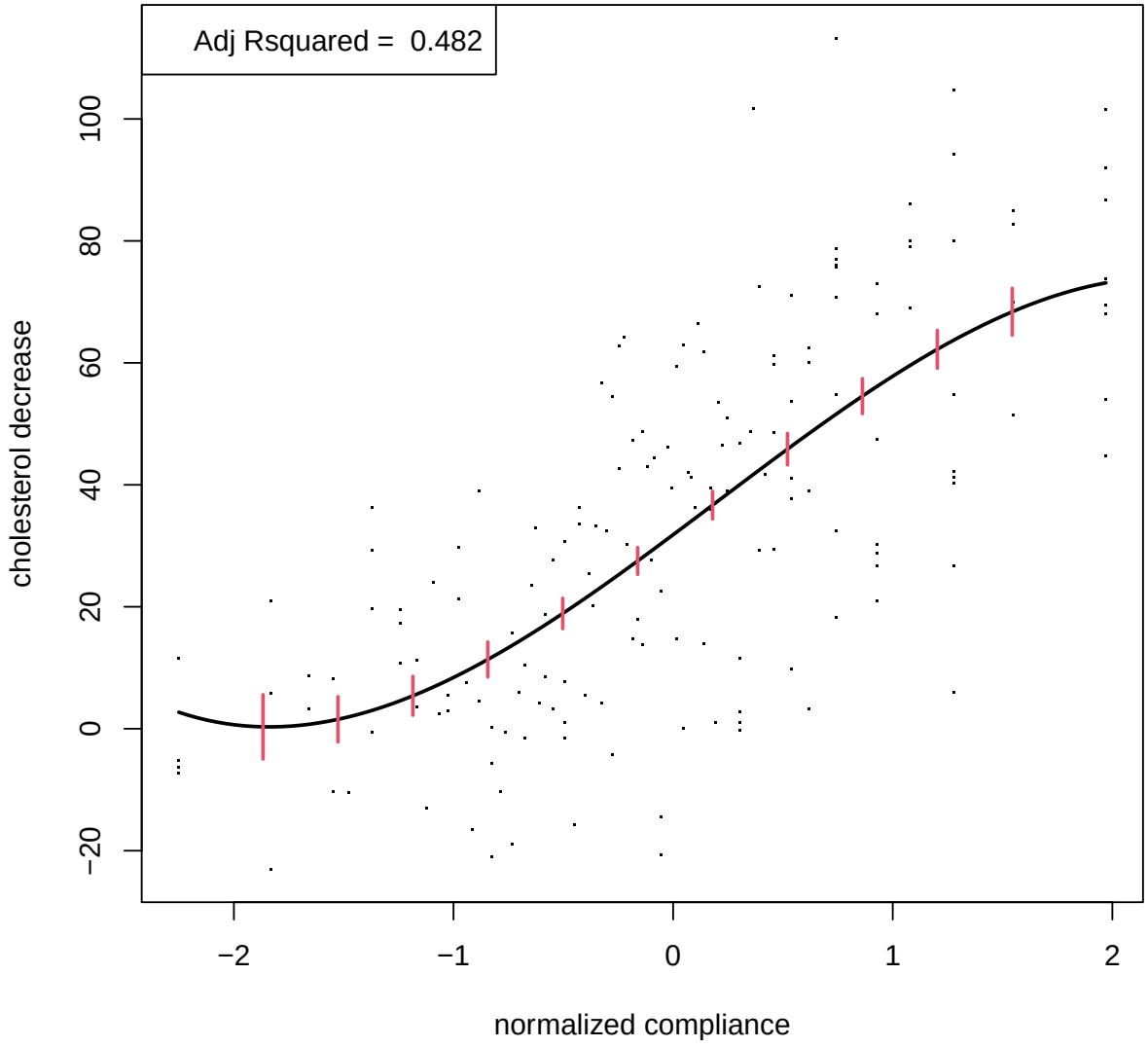


Figure 2. On left, a surface depicting Newton's second law of motion, acceleration = force/mass; on right, a noisy version.

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



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

- The figure shows a small example, taken from a larger dataset in Efron and Feldman (1991): $n = 164$ male doctors volunteered to take the cholesterol-lowering drug cholostyramine.
- Two numbers were recorded for each doctor, x_i = normalized compliance and y_i = observed cholesterol decrease.
- Compliance, the proportion of the intended dose actually taken, ranged from 0% to 100%, −2.25 to 1.97 on the normalized scale, and of course it was hoped to see larger cholesterol decreases for the better compliers.

- 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 $\mathcal{S}$. That is the estimation side of the story.
- As far as attribution is concerned, 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.

`birthwt` data

- R package `MASS`
- The birthwt data frame has 189 rows and 10 columns.
- The data were collected at Baystate Medical Center, Springfield, Mass during 1986.

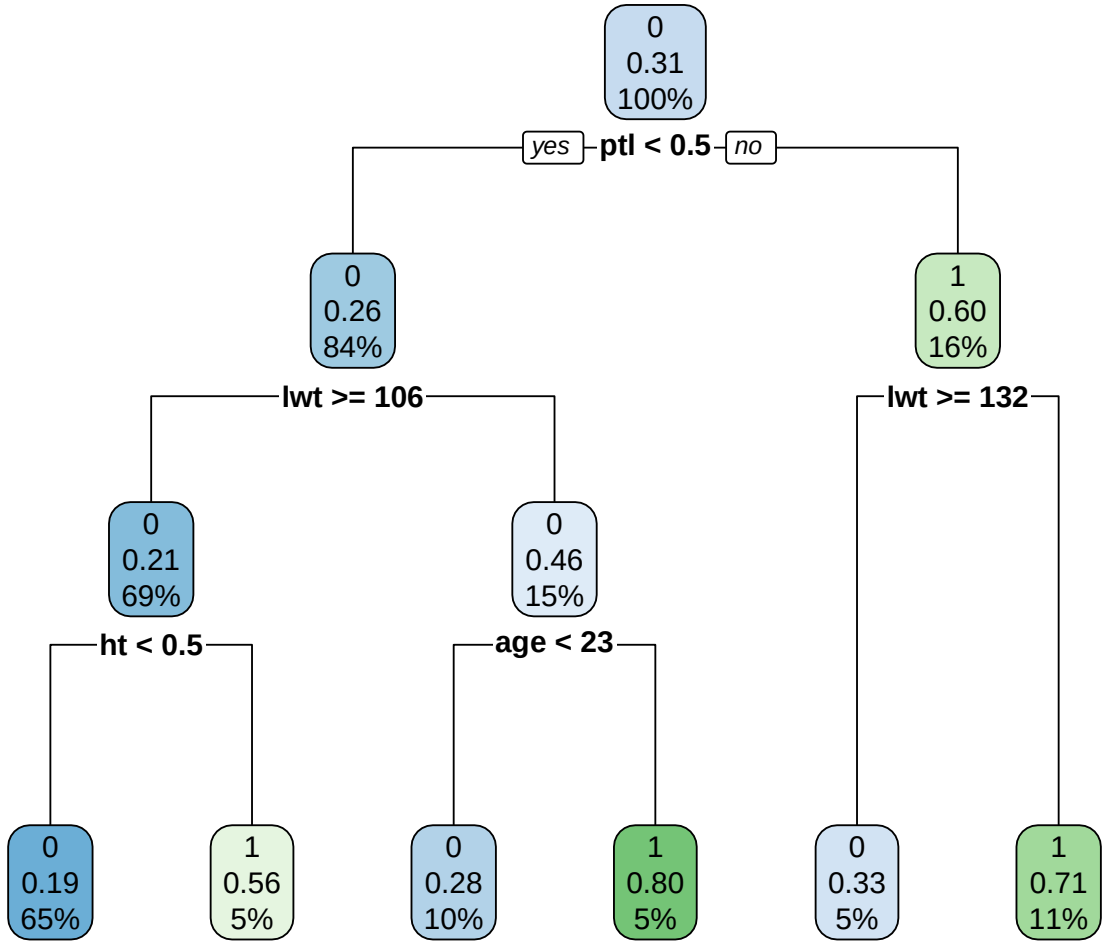
- Another mainstay of traditional methodology is logistic regression.
- The dataset concerns the Risk Factors Associated with Low Infant Birth Weight: $n = 189$ babies, 59 with birth weight less than 2.5 kg and 130 with more than 2.5 kg.
- Eight covariates were measured at entry: mother's age in years, mother's weight in pounds at last menstrual period, body weight, etc., so x_i was 8-dimensional, while y_i equaled 0 or 1
- This is a surface plus noise model, with a linear logistic surface and Bernoulli noise.

	term	estimate	std.error	p.value	
1	(Intercept)	1.07	1.27	0.40	
2	age	-0.04	0.04	0.31	
3	lwt	-0.02	0.01	0.02	*
4	raceblack	1.12	0.54	0.04	*
5	raceother	0.67	0.47	0.16	
6	smoke	0.75	0.43	0.08	
7	ptl	-1.66	0.90	0.07	
8	ht	1.93	0.73	0.01	**
9	ui	0.80	0.48	0.09	
10	ftv1	-0.52	0.49	0.29	
11	ftv2+	0.10	0.46	0.83	
12	ptd	3.41	1.22	0.01	**

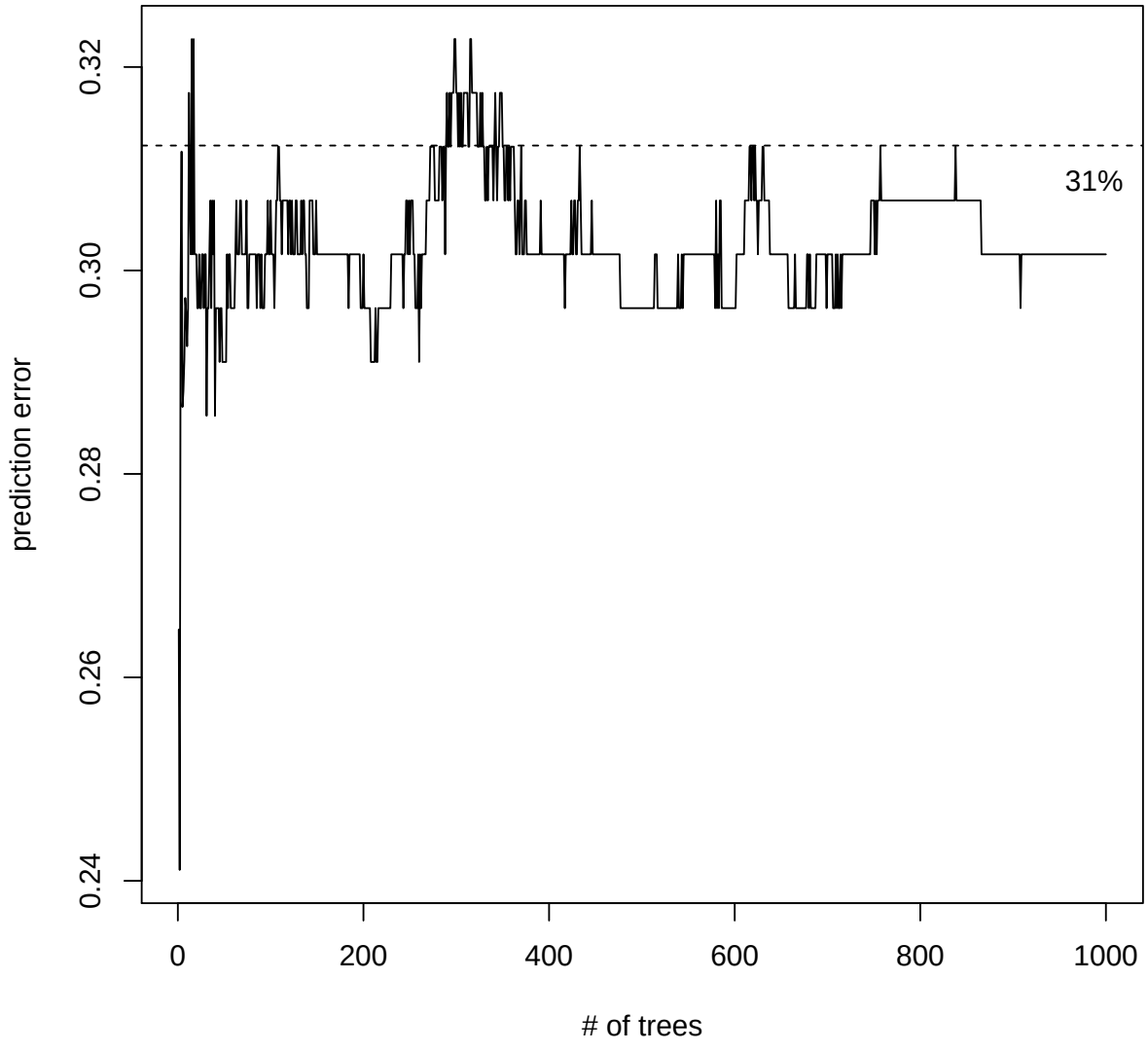
3. The Pure Prediction Algorithms

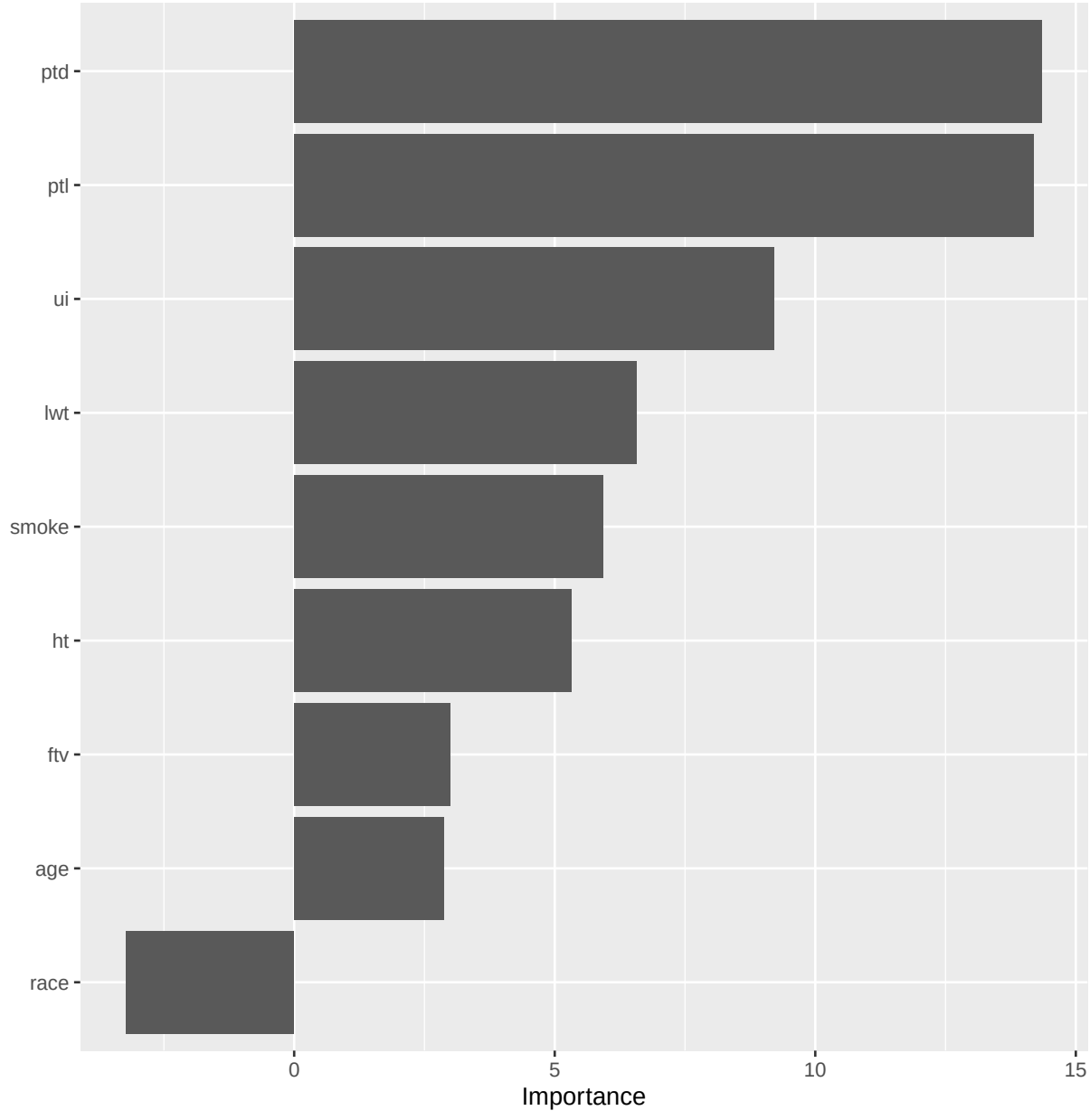
- Random Forests, Boosting, Deep Learning, etc.
- Data
$$\mathbf{d} = \{(x_i, y_i), i = 1, \ldots, n\}$$
- Prediction rule $f(x, \mathbf{d})$
- New $(x, ?)$ gives $\hat{y} = f(x, \mathbf{d})$
- Strategy: Go directly for high predictive accuracy; forget (mostly) about surface + noise

CART


0
0.31
100%
yes
ptl < 0.5
no
0
0.26
84%
1
0.60
16%
lwt >= 106
lwt >= 132
0
0.21
69%
0
0.46
15%
ht < 0.5
age < 23
0
0.19
65%
1
0.56
5%
0
0.28
10%
1
0.80
5%
0
0.33
5%
1
0.71
11%

Random forest




ptd
ptl
ui
lwt
smoke
ht
ftv
age
race
0
5
10
15
Importance

Apparent error rate (training error)

$$\widehat{\text{err}} = \#\{f(x_i, \mathbf{d}) \neq y_i\}/n$$

	model	error rate
1	rand_forest	0.222
2	logistic_reg	0.243
3	decision_tree	0.228

True error rate

$$E(f(X, \mathbf{d}) \neq Y)$$

where (X, Y) is a random draw from whatever probability distribution gave the (x_i, y_i) pairs in $\mathbf{d}$;

Estimated by 10-fold cross-validated error rate

	model	mean	n	std_err
1	rand_forest	0.312	10	0.04
2	logistic_reg	0.313	10	0.03
3	decision_tree	0.365	10	0.04

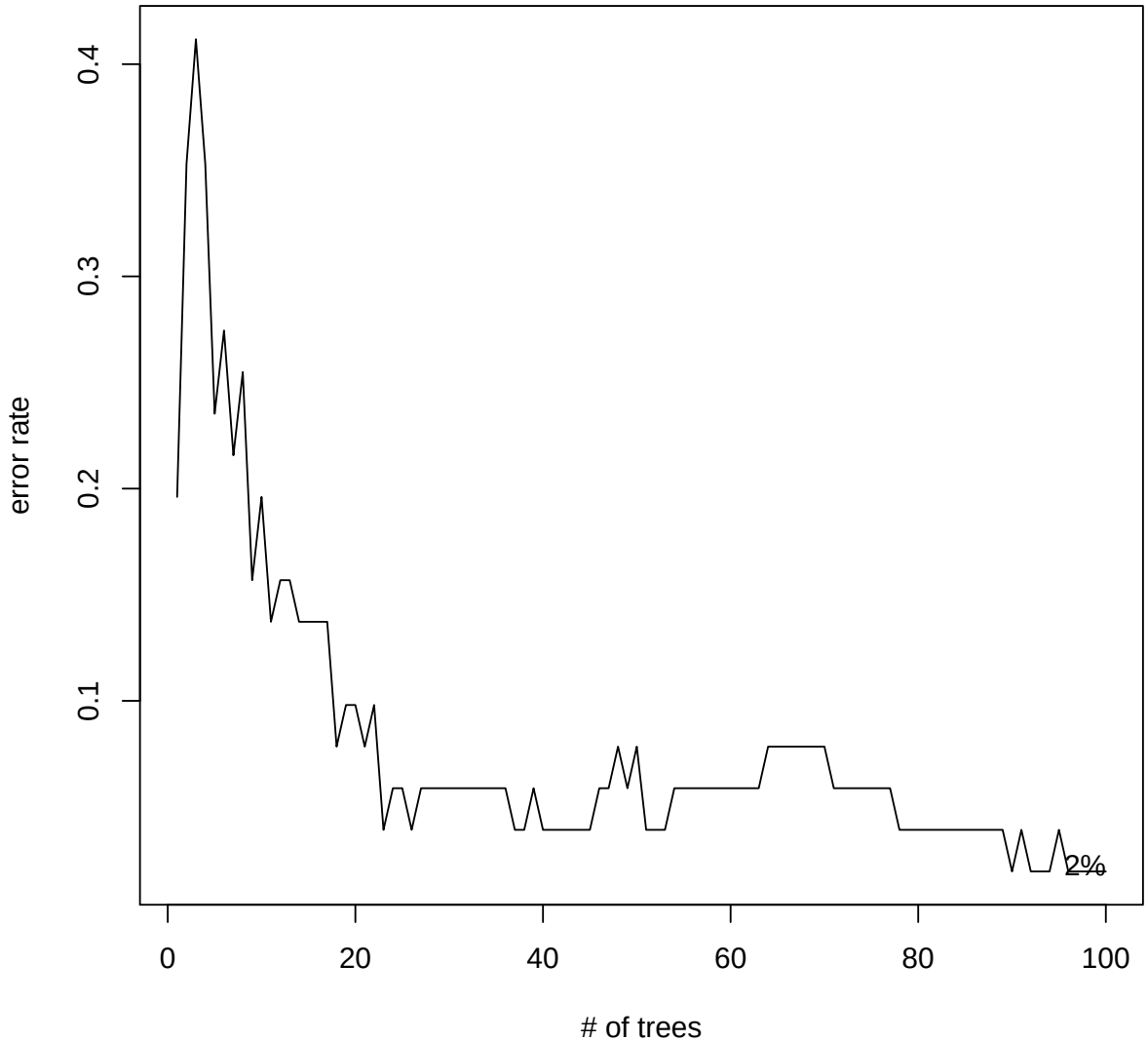
4. A Microarray Prediction Problem

The Prostate Cancer Microarray Study

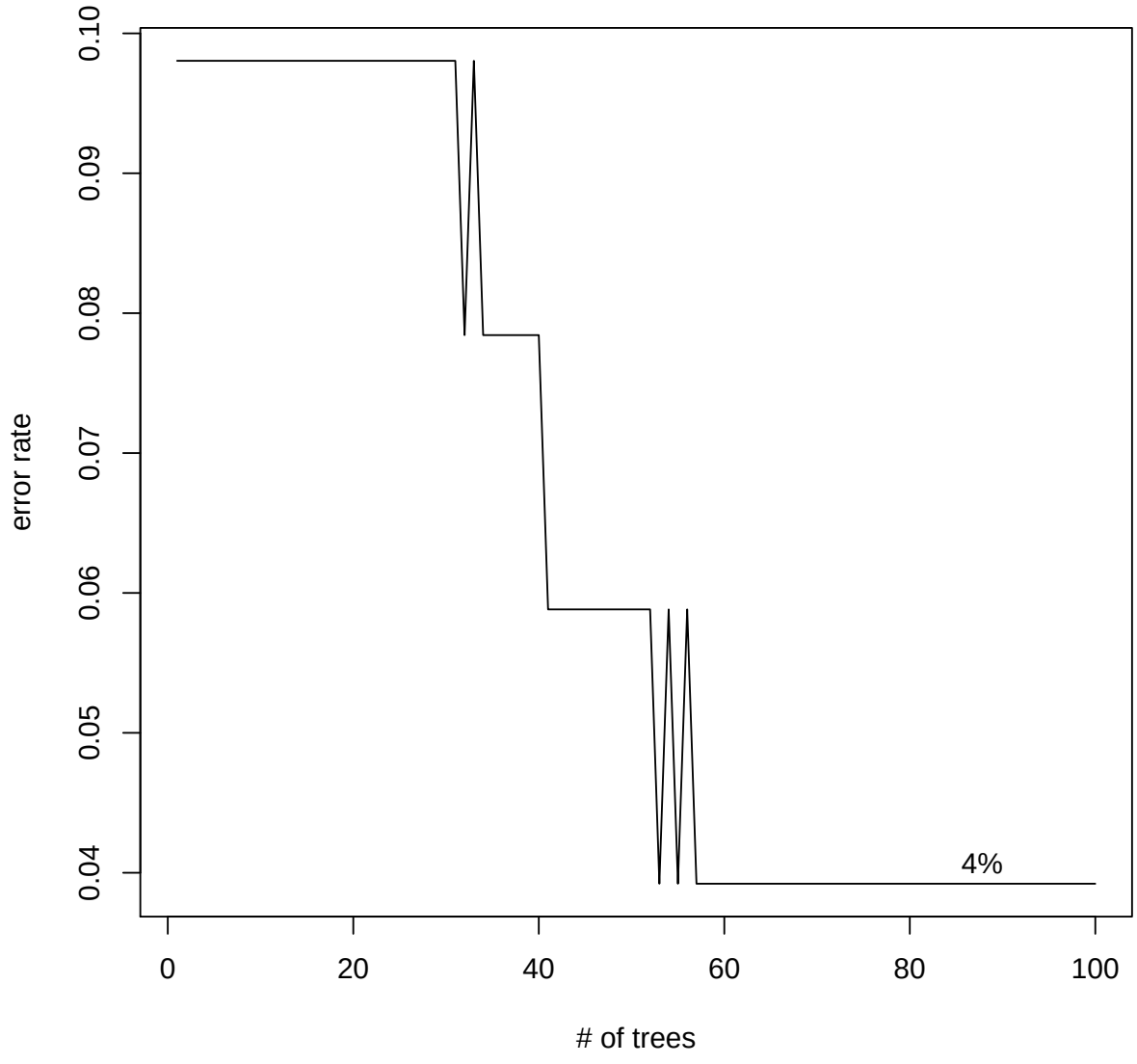
- https://hastie.su.domains/CASI_files/DATA/prostate.html
- $n = 100$ men: 50 prostate cancer, 50 normal controls
- For each man measure activity of $p = 6033$ genes
- Data set $\mathbf{d}$ is 100×6033 matrix ("wide")
- Wanted: Prediction rule $f(x, \mathbf{d})$ that inputs new 6033-vector x and outputs $\hat{y}$ correctly predicting cancer/normal

Random forest

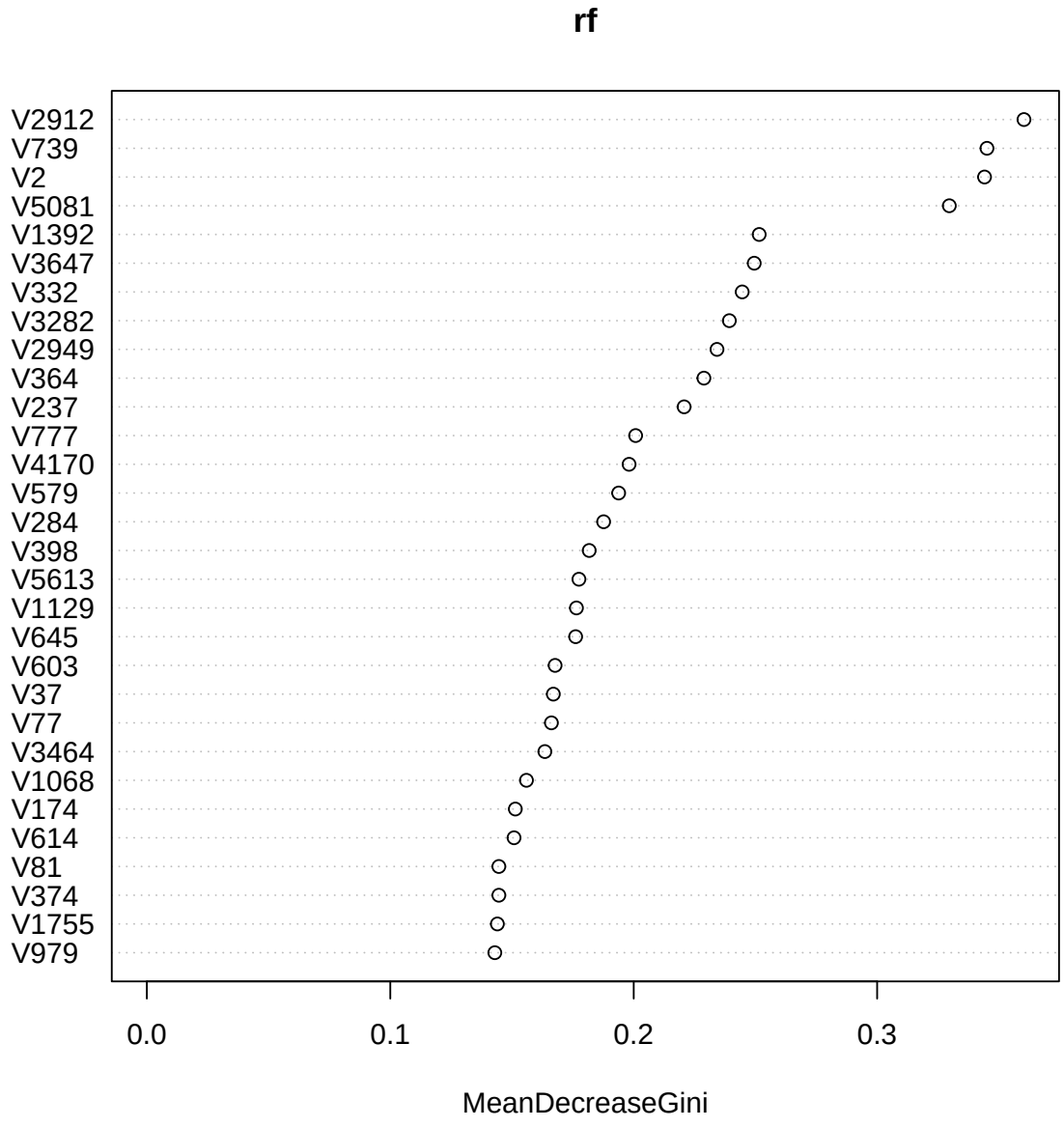
- Randomly divide the 102 subjects into:
  - training set of 51 subjects (25 + 25)
  - test set of 51 subjects (25 + 25)
- Run R program `randomForest` on the training set
- Use its rule $f(x_i, \mathbf{d})$ on the test set and see how many errors it makes


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

Boosting


error rate
0.04
0.05
0.06
0.07
0.08
0.09
0.10
0
20
40
60
80
100
# of trees
4%

5. Advantages and Disadvantages of Prediction


rf
V2912
V739
V2
V5081
V1392
V3647
V332
V3282
V2949
V364
V237
V777
V4170
V579
V284
V398
V5613
V1129
V645
V603
V37
V77
V3464
V1068
V174
V614
V81
V374
V1755
V979
0.0
0.1
0.2
0.3
MeanDecreaseGini