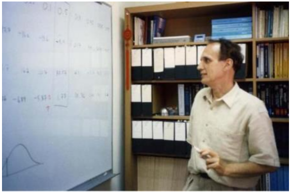
Prediction, Estimation, and Attribution

Statistical Learning
CLAMSES - University of Milano-Bicocca

Aldo Solari



Bradley Efron working in his classic office, circa 1996.

References

This material is very much based on:

- 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**.

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.

638 B. EFRON

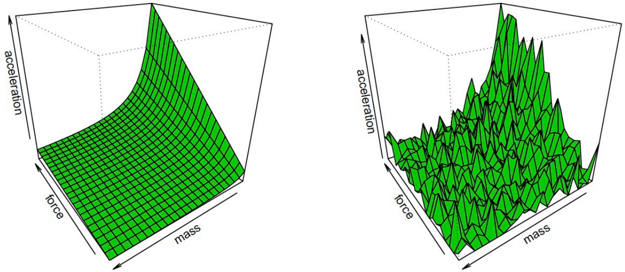
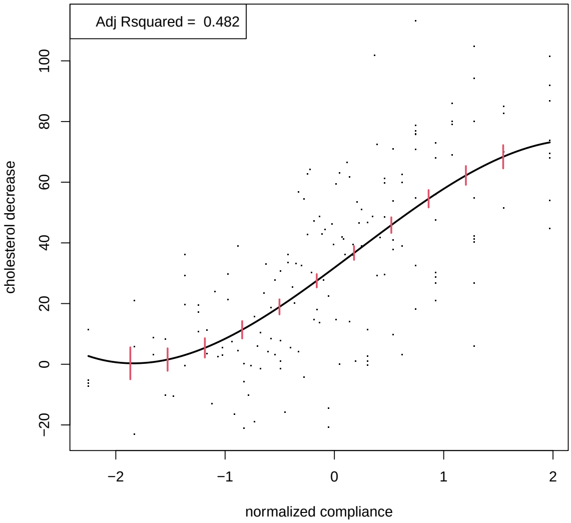


Figure 2. On left, a surface depicting Newton's second law of motion, acceleration = force/mass; on right, a noisy version.

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.

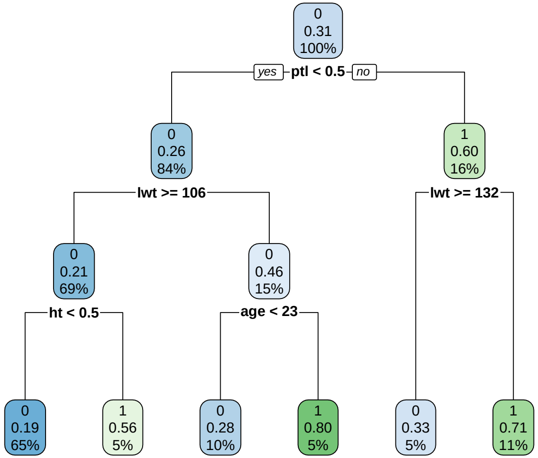
- Another mainstay of traditional methodology is logistic regression.
- The dataset concerns the Risk Factors Associated with Low Infant Birth Weight: $n = 189$ babies, 59 with birth weight less than 2.5 kg and 130 with more than 2.5 kg.
- Eight covariates were measured at entry: mother's age in years, 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	**

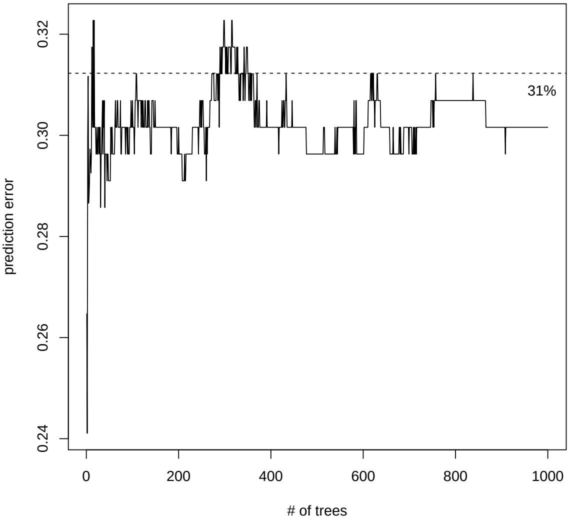
3. The Pure Prediction Algorithms

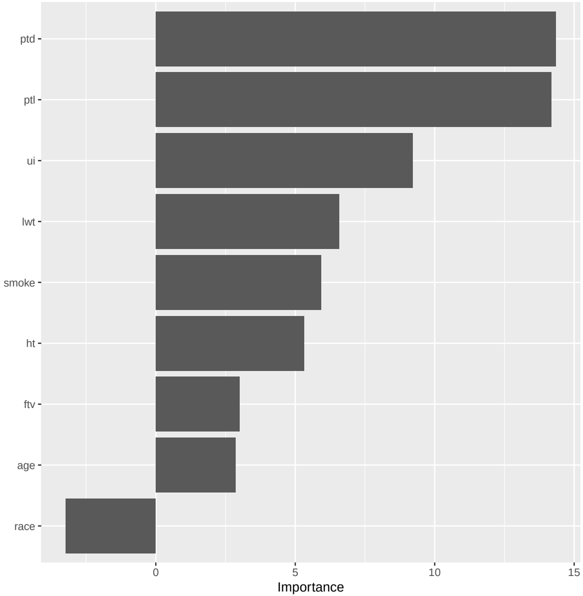
- Random Forests, Boosting, Deep Learning, etc.
- Data
$$\mathbf{d} = \{(x_i, y_i), i = 1, \ldots, n\}$$
- Prediction rule $f(x, \mathbf{d})$
- New $(x, ?)$ gives $\hat{y} = f(x, \mathbf{d})$
- Strategy: Go directly for high predictive accuracy; forget (mostly) about surface + noise

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

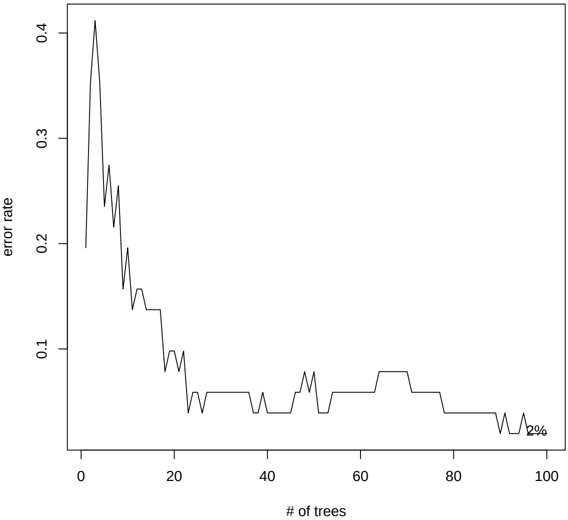
4. A Microarray Prediction Problem

The Prostate Cancer Microarray Study

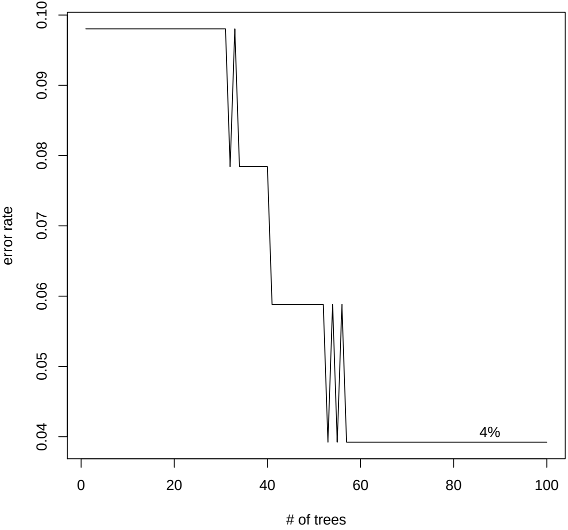
- https://hastie.su.domains/CASI_files/DATA/prostate.html
- $n = 100$ men: 50 prostate cancer, 50 normal controls
- For each man measure activity of $p = 6033$ genes
- Data set $\mathbf{d}$ is 100×6033 matrix ("wide")
- Wanted: Prediction rule $f(x, \mathbf{d})$ that inputs new 6033-vector x and outputs $\hat{y}$ 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
0.4
0.3
0.2
0.1
0
20
40
60
80
100
# of trees
2%

Boosting



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

5. Advantages and Disadvantages of Prediction

Prediction is Easier than Estimation

- Observe
  - $x_1, \ldots, x_{25} \overset{\text{ind}}{\sim} \mathcal{N}(\mu, 1)$
  - $\bar{x}$ sample mean, $\tilde{x}$ sample median,
- Estimation

$$E\{(\tilde{x} - \mu)^2\}/E\{(\bar{x} - \mu)^2\} = 1.57$$

- Wish to predict new $X \sim \mathcal{N}(\mu, 1)$
- Prediction

$$E\{(\tilde{x} - X)^2\}/E\{(\bar{x} - X)^2\} = 1.02$$

- The reason is that most of the prediction error comes from the variability of X, which neither $\bar{x}$ or $\tilde{x}$ can cure
- Prediction is easier than estimation, at least in the sense of being more forgiving. This allows for the use of inefficient estimators like the gbm stumps

Prediction is Easier than Attribution

- Microarray study involving n subjects, $n/2$ healthy controls and $n/2$ sick patients
- Each subject provides a vector of measurements on N genes $\mathbf{x} = (x_1, \ldots, x_N)^t$ with

$$X_j \overset{\text{ind}}{\sim} \mathcal{N}(\pm\delta_j/2c, 1)$$

  where $c = \sqrt{(n/4)}$ "plus" for the sick and "minus" for the healthy; δ_j is the effect size for gene j.
- N_0 genes are null i.e. $\delta_j = 0$
- a small number N_1 of genes are non-null and have $\delta_j = \Delta$
- A new person arrives and produces a microarray of measurements $\mathbf{x} = (x_1, \ldots, x_N)^t$ but without us knowing the person's healthy/sick status; that is, without knowledge of the $\pm$ value
- Question: How small can N_1/N_0 get before prediction becomes impossible?

Prediction is Easier than Attribution

- Asymptotically as $N_0 \to \infty$, accurate prediction is possible if

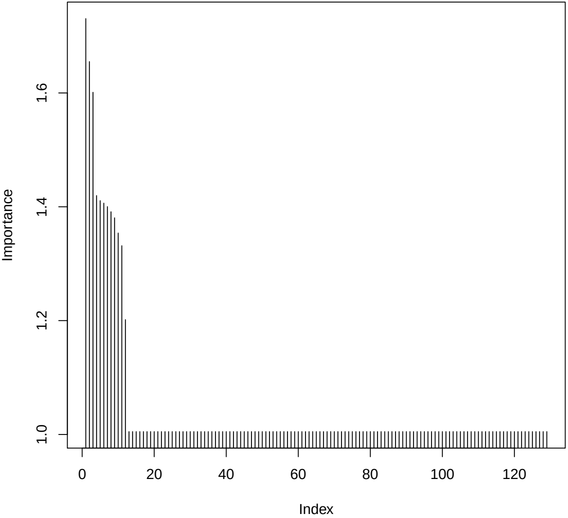
$$N_1 = O(N_0^{1/2})$$

- Effective attribution requires

$$N_1 = O(N_0)$$

- In terms of “needles in haystacks”, attribution needs an order of magnitude more needles than prediction.
- The three main regression categories can usefully be arranged in order

prediction estimation attribution

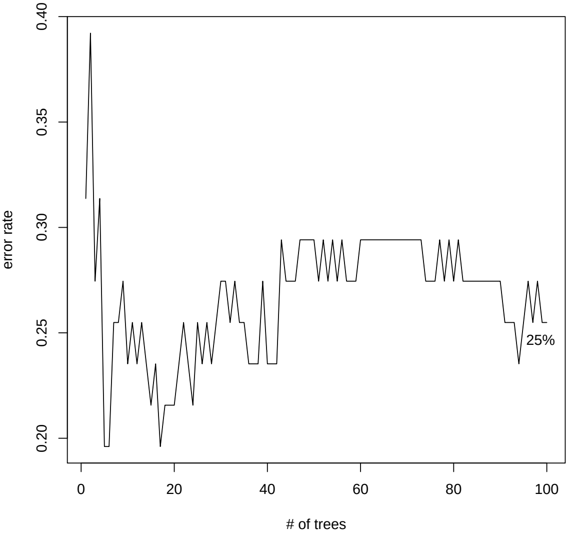

Importance
1.0
1.2
1.4
1.6
0
20
40
60
80
100
120
Index

- 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.

6. The Training/Test Set Paradigm

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)


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

Hypothetical microarray study

- $n = 400$ subjects participate in the study, arriving one per day in alternation between Treatment and Control
- Each subject is measured on a microarray of $p = 200$ genes
- The $400 \times \; 200$ data matrix $\mathbf{x}$ has independent normal entries

$$X_{ij} \overset{\text{ind}}{\sim} \mathcal{N}(\mu_{ij}, 1)$$

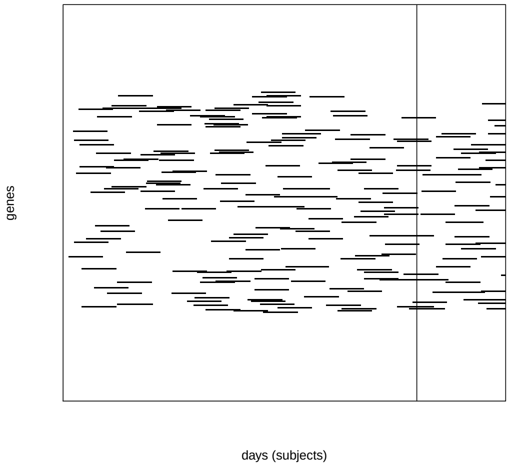
- Most of the μ_{ij} are null, $\mu_{ij} = 0$, but occasionally a gene will have an active episode of 30 days during which

$$\mu_{ij} = 2 \quad \text{for Treatment} \quad \mu_{ij} = -2 \quad \text{for Control}$$

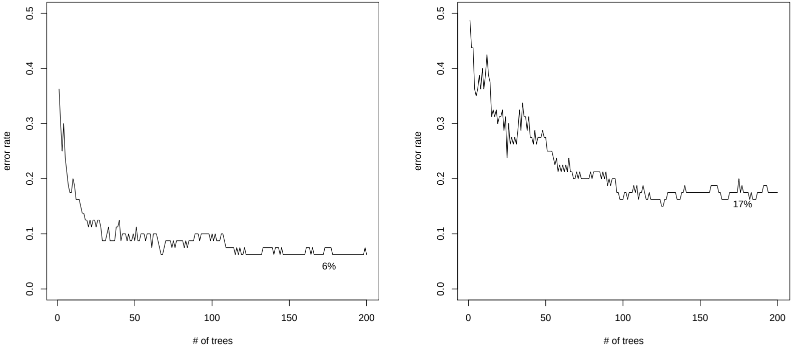
  for the entire episode, or

$$\mu_{ij} = -2 \quad \text{for Treatment} \quad \mu_{ij} = 2 \quad \text{for Control}$$

  for the entire episode.
- Each gene has expected number of episodes equal 1



Black line segments indicate active episodes in the hypothetical microarray study. (Matrix transposed for typographical convenience.)



randomForest prediction applied to the hypothetical microarray study microarray study. Left panel: Test set of size 80, selected randomly from 400 days; Right panel: Test set days 321-400

- From any one day's measurements it is possible to predict Treatment or Control from the active episode responses on nearby days
- This works for the random training/test division, where most of the test days will be intermixed with training days.
- Not so for the early/late division, where most of the test days are far removed from training set episodes.
- To put it another way, prediction is easier for interpolation than extrapolation

Replicability

Year (study) 1: $n = 812$, $p = 11$ (selected from an initial list of 81)

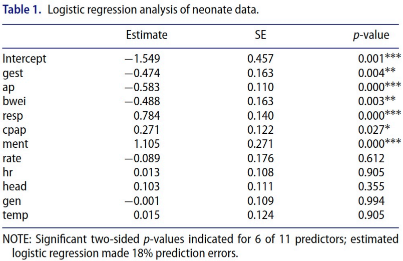


Table 1. Logistic regression analysis of neonate data.

	Estimate	SE	*p*-value
Intercept	−1.549	0.457	0.001***
gest	−0.474	0.163	0.004**
ap	−0.583	0.110	0.000***
bwei	−0.488	0.163	0.003**
resp	0.784	0.140	0.000***
cpap	0.271	0.122	0.027*
ment	1.105	0.271	0.000***
rate	−0.089	0.176	0.612
hr	0.013	0.108	0.905
head	0.103	0.111	0.355
gen	−0.001	0.109	0.994
temp	0.015	0.124	0.905

NOTE: Significant two-sided *p*-values indicated for 6 of 11 predictors; estimated logistic regression made 18% prediction errors.

Year (study) 2: $n = 246$, $p = 11$

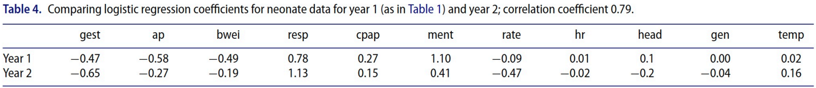


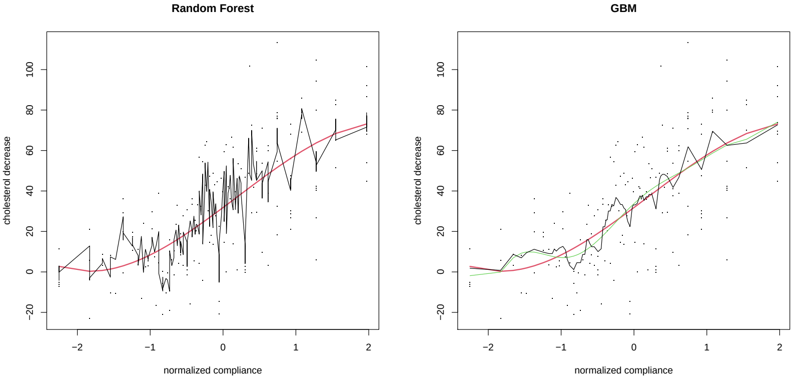
Table 4. Comparing logistic regression coefficients for neonate data for year 1 (as in Table 1) and year 2; correlation coefficient 0.79.

	gest	ap	bwei	resp	cpap	ment	rate	hr	head	gen	temp
Year 1	−0.47	−0.58	−0.49	0.78	0.27	1.10	−0.09	0.01	0.1	0.00	0.02
Year 2	−0.65	−0.27	−0.19	1.13	0.15	0.41	−0.47	−0.02	−0.2	−0.04	0.16

7. Smoothness

- Estimation and Attribution: seek long-lasting scientific truths
    - physics
    - astronomy
    - medicine
    - economics?
- Prediction algorithms: truths and ephemeral relationships
    - credit scores
    - movie recommendations
    - image recognition
- Estimation and Attribution: theoretical optimality (MLE, Neyman-Pearson)
- Prediction: training-test performance
- Nature: rough or smooth?

- 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 $\mathbf{x}$ 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.

8. A Comparison Checklist

	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.	$\mathbf{x}$ $n \times p$ 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)

9. Traditional Methods in the Wide Data Era

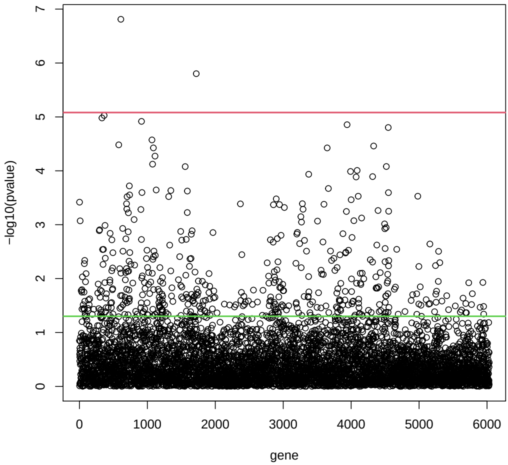
Estimation and Attribution in the Wide-Data Era

- 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

$$p_i \leq 0.05/6033$$

$$\begin{aligned} \Pr(\text{Type I error} > 0) &= \Pr\Big(\bigcup_{i \in \mathcal{T}} \{p_i \leq \alpha/p\}\Big) \\ &\leq \sum_{i \in \mathcal{T}} \mathrm{P}(p_i \leq \alpha/p) \leq |\mathcal{T}| \frac{\alpha}{p} \leq \alpha \end{aligned}$$

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

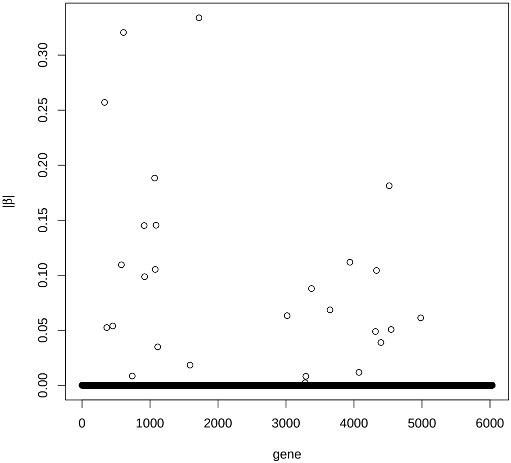


−log10(pvalue)
0
1
2
3
4
5
6
7
gene
0
1000
2000
3000
4000
5000
6000

- 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

10. Two Hopeful Trends

- 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.