Discussion of
Estimation and Accuracy After Model Selection
By Bradley Efron

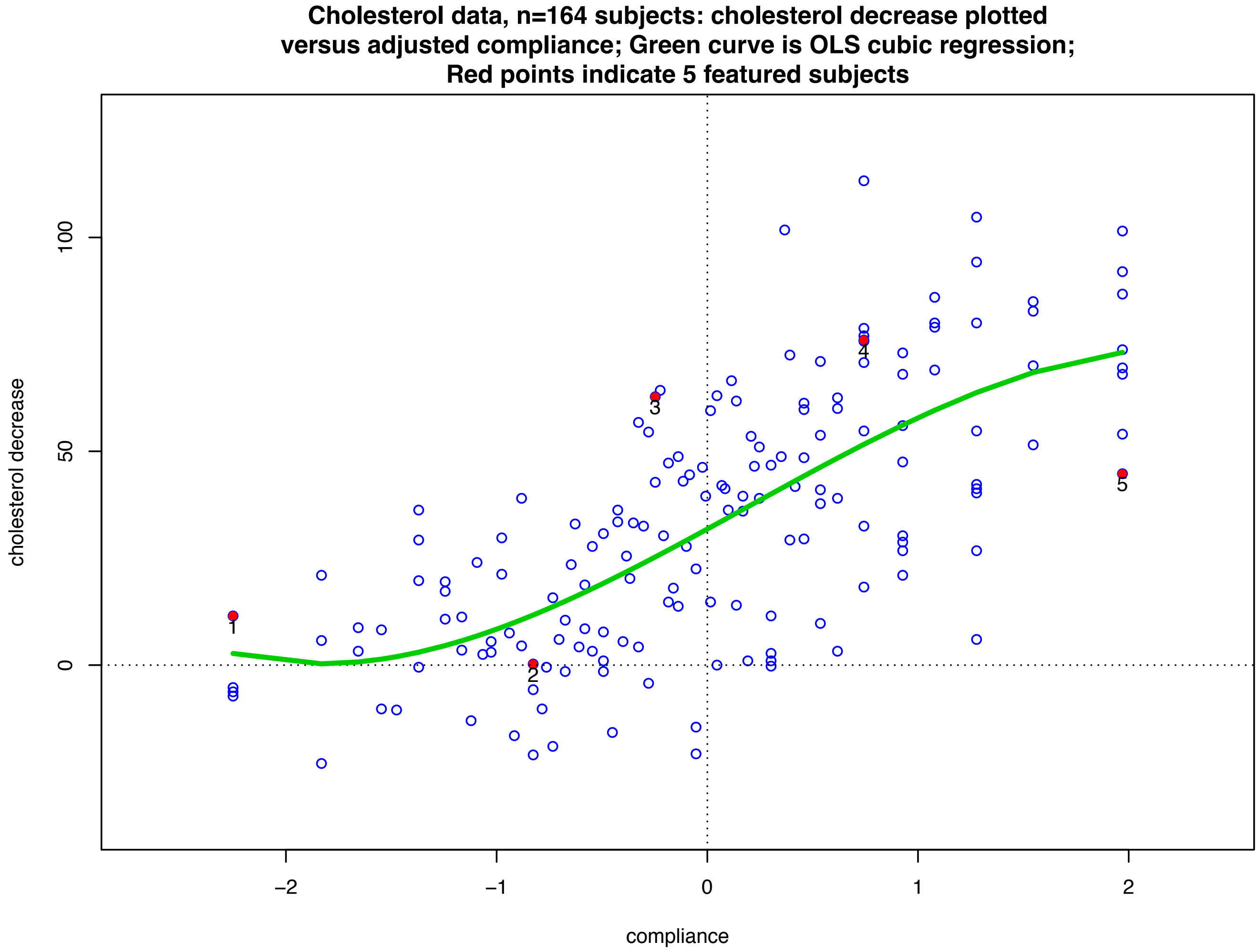
Discussant: Lawrence D Brown*

Statistics Department, Wharton, Univ. of Penn.
lbrown@wharton.upenn.edu

JSM, Boston, Aug. 4, 2014

* Joint author: Dan McCarthy (Univ. of Penn.)

Discussion is about analysis of Cholesterol Data



Focus on:

- Initial comment about **Po**st **S**election **I**nference
- The "bagged" estimator with C_p selection, vs
  a SURE mixture of polynomials.
- Confidence intervals for the predictive mean via
  a direct double bootstrap

Post **S**election **I**nference

Efron points out it is bad practice to:

(a) look at data
(b) choose model
(c) fit estimates using chosen model
(d) analyze [get CIs] as if pre-chosen

as Efron notes Berk, …, Zhao (2013), *Ann Stat,* 802-813 [“PoSI”] make a similar point.

But the problem considered there is different from that here:

Differences

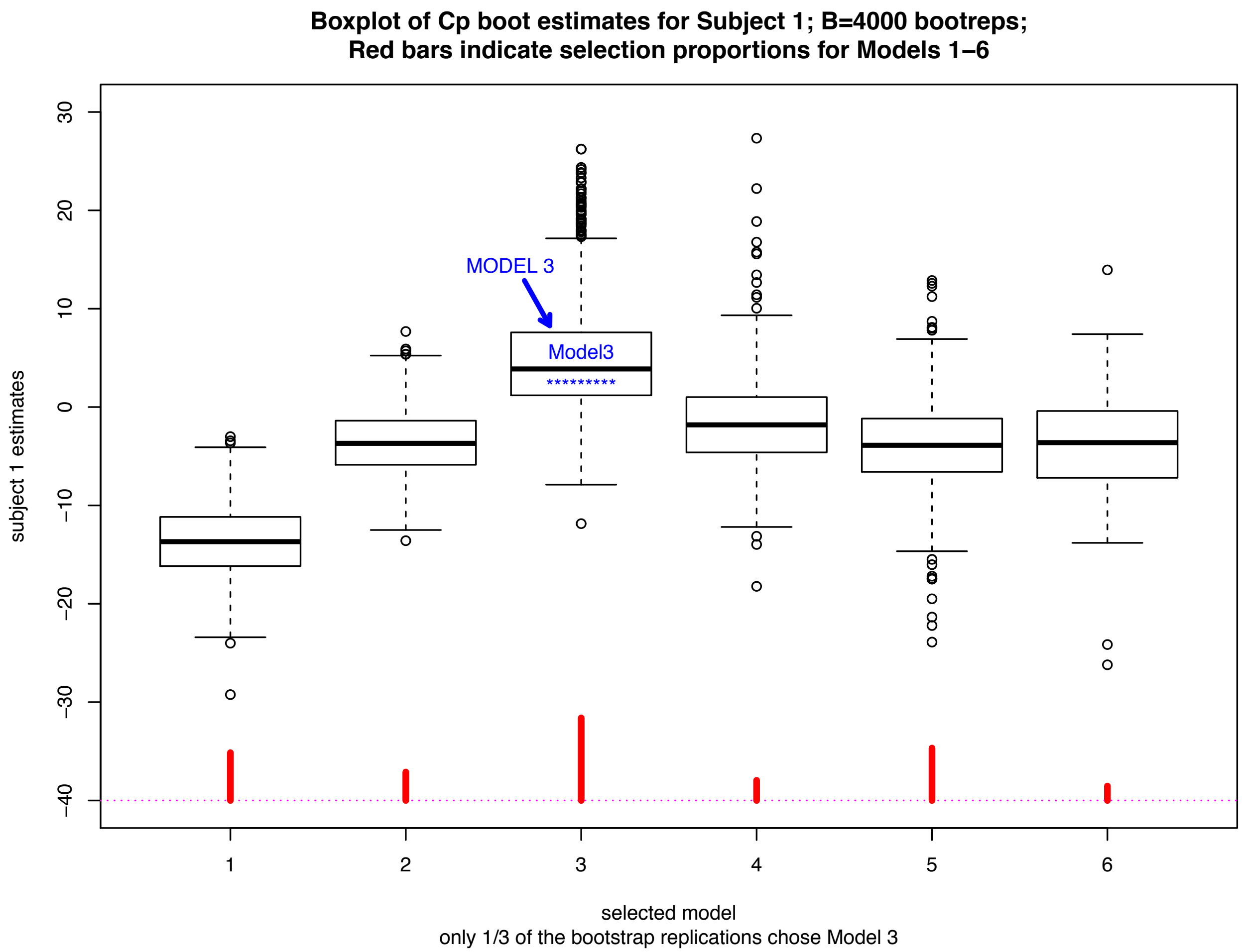
	"PoSI" paper	Efron
1	Inference about slopes	Inference about $E(Y\|x_k)$
2	Model algorithm not pre-specified	Model via C_p
3	Fixed design	Random design for covariates
4	Conventional assumption on residuals	No assumption on residuals

#1 is important to POSI since the error of slope estimates does not depend on parameters

However, **#2** for Efron allows narrower CIs. & It's suitable for bootstrapping.

#3 for Efron allows for the paired bootstrap.

So Efron implements a pairs-bootstrap and gets:



Boxplot of Cp boot estimates for Subject 1; B=4000 bootreps;
Red bars indicate selection proportions for Models 1–6
subject 1 estimates
30
20
10
0
−10
−20
−30
−40
MODEL 3
Model3
****
1
2
3
4
5
6
selected model
only 1/3 of the bootstrap replications chose Model 3

Embarrassing because:

1. Raises issue of what is “true” target for the bootstrap.

2. Suggests Efron’s estimate from the data may be badly biased.

3. Calls into question the integrity of C_p selection as basis for estimation + CI.

SO

4. Efron recommends “bagging” as a way to produce a more stable and smoother estimator, which may yield more satisfactory estimate and bootstrap CIs.

My two main topics

A. An alternate method to directly produce an estimate that is an average of polynomials.

B. An alternate bootstrap methodology. (Applied to the bagged estimator, but could also apply to the estimator in A.)

Alternate Methodology, The SURE-weighted average of polynomials

- Let $\vec{v}_1, \vec{v}_2, \ldots$ be the sequentially orthogonalized versions of $\vec{x}, \overrightarrow{x^2}, \ldots$. (ie, $\vec{v}_1 \perp \mathbf{1}$, $\vec{v}_2 \perp 1 \& \vec{v}_1$, …)
- A weighted average of L.S. polynomials is

$$\hat{f}(x) = \hat{\gamma}_0 + \omega_1 \hat{\gamma}_1 + \ldots + \omega_p \hat{\gamma}_p \ni$$

(*) $$1 = \omega_1 \geq \omega_2 \geq \ldots \geq \omega_p$$

- Then

$$\text{SURE} = \text{popSSE} + 2\sum\left(\omega_j - 1\right)\hat{\sigma}^2 + \sum\left(\omega_j - 1\right)^2 \hat{\gamma}_j^2 .$$

- Minimize this subject to the monotonicity (*).
- This yields estimates for the weights

- The unconstrained minimizer of SURE is

$$\omega_{j;\text{uncon}} = 1 - \left(1 \wedge \left(1 \big/ \hat{\gamma}_j^2\right)\right)$$

- but the constraint (*) may require a PAV operation that pools adjacent coordinates and produces J-S shrinkage among them.

- The result is very similar to the bagged estimator, but not exactly the same. (It's nearly indistinguishable on the very benign Cholesterol data.)

The Double Bootstrap

- This is similar to what is suggested in DiCiccio and Efron (1996) *Stat Sci*, and elsewhere
- But without any BCa/ABC step.
- We find it to work well here, and in (the few) other simple and multiple regression examples we've so far tried it on.
- Here is a schematic:

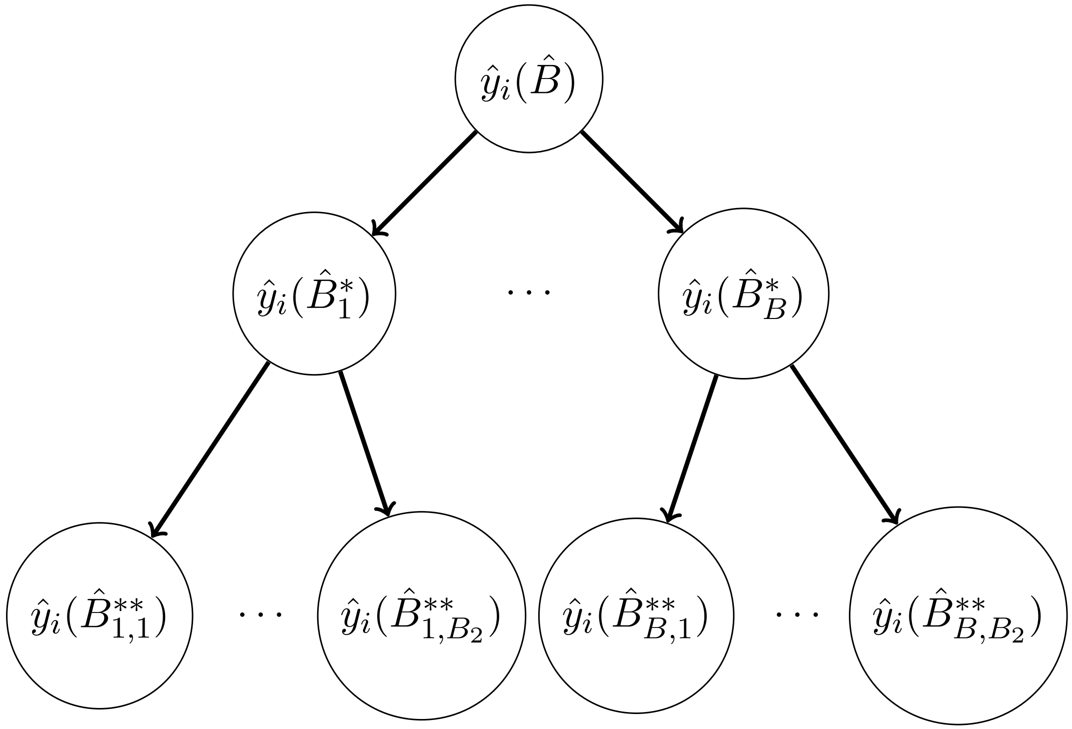


Figure 1: Two Levels of Bootstrap Conditional Mean Estimates for Observation i

$\hat{y}\left(\hat{B}\right)$ denotes the original bagged estimator.

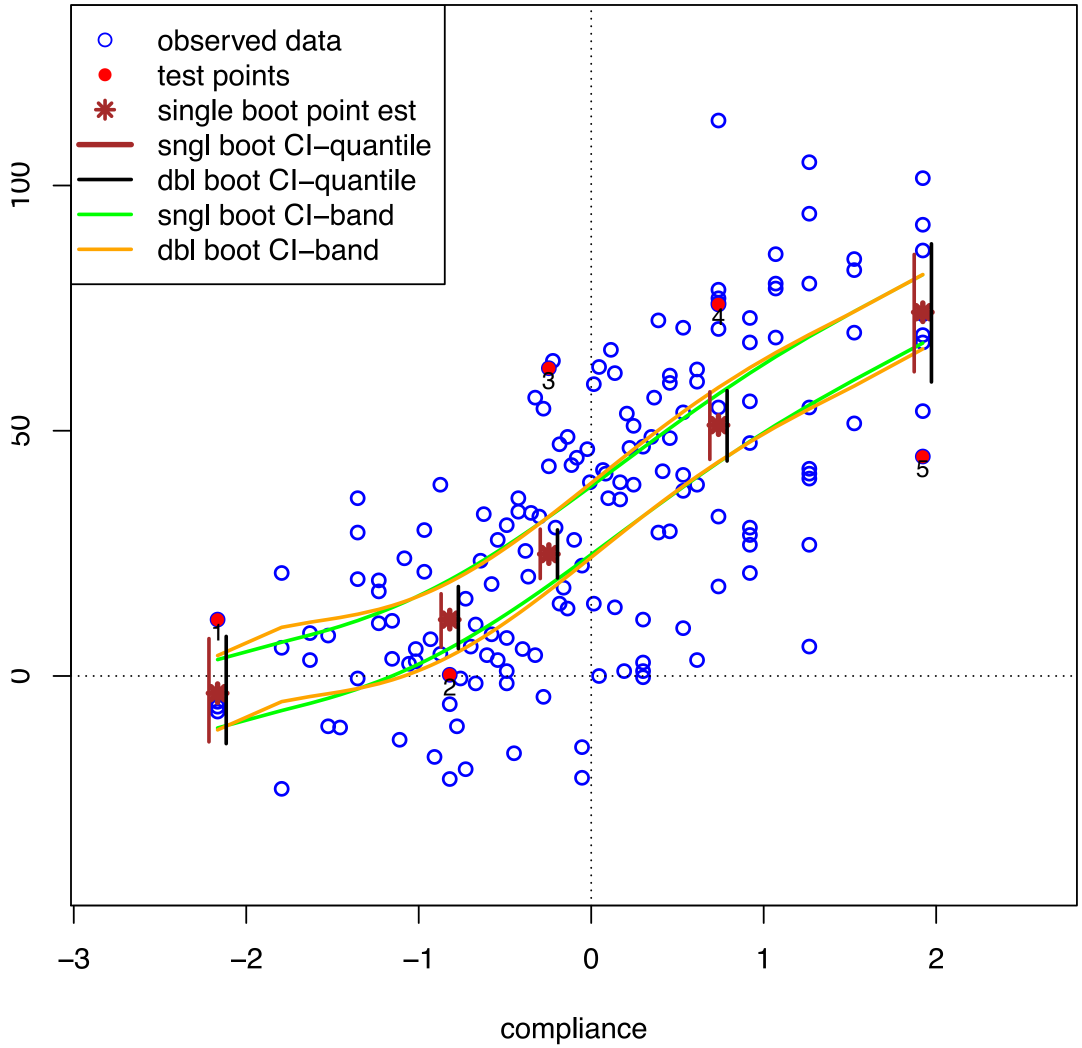
The other levels are paired-bootstraps followed by bagging estimators. Second bootstrap results are used to calibrate the first:

- For equal- tail intervals the first level bootstrap yields CIs of the form $L_{\beta/2}, U_{1-\beta/2}$ that putatively cover with probability $1-\beta$.
- The second level calibrates (adjusts) the value of β so that the actual coverage is estimated to be the desired $1-\alpha$.
- The resulting CIs are equal-tail but need not be symmetric about the point estimate.
- Two types of 95% CIs were computed

(1) CIs for estimates of $E(Y|x_k)$ at Efron's 5 points

(2) Parallel bands that have marginal probability 95% of covering E(Y|X).

Cholesterol data, n=164 subjects: cholesterol decrease plotted versus adjusted compliance; Confidence Intervals Added; Red points indicate 5 featured subjects



Notes:

- What's labeled "single boot point estimate" is actually the bagging estimate from the original data.
- The differences here between the CIs is pretty small. This is attributable to the well-behaved data. For less benign data the differences can be much more notable.
- Typically (but not always) the double boot CIs are wider than single boot ones. That's so here, but not too noticeable. (But look at the right-most of the 5 points.)
- The double boot routine is computationally intensive in its own right and <u>much</u> more so because the bagged estimator itself requires 500 bootstrap samples each time it is computed; for the SURE estimator the routine would be quite fast to compute (minutes instead of 4 hours on a parallel array).