Estimation and Accuracy after Model Selection

Yongchan Kwon

Department of Statistics, Seoul National University, Seoul, Korea

February 26, 2016

Contents

1. Introduction
2. Nonparametric bootstrap smoothing
3. Take a break
4. Accuracy of the smoothed bootstrap estimates
5. Parametric bootstrap smoothing
6. Discussions

Motivating example

- $n = 164$ men took Cholestyramine for ~ 7 years
- $x =$ compliance measure (normalized)
- $y =$ cholesterol decrease
- Regression y on x ? Wish to estimate $\mu_j = E[y_j \mid x_j]$ for $j = 1, \ldots, n$

Motivating example

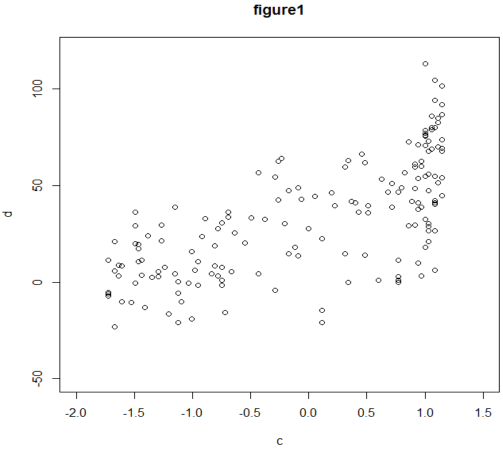


Figure: Cholesterol data, n=164 subjects: cholesterol decrease plotted versus normalized compliance

Motivating example

- Regression Model: $y = X\beta + e$, $[e_i \sim (0, \sigma^2)]$
- C_p Criterion: $\|y - X\hat{\beta}\|^2 + 2m\sigma^2$
  $\hat{\beta}$ =OLS estimate, m ="degree of freedom"
- Model Selection: from possible models $M_1, M_2, M_3, \ldots$, choose the one minimizing C_p.
- Then use OLS estimate from chosen model. (Assume model selection procedure is known)

Motivating example

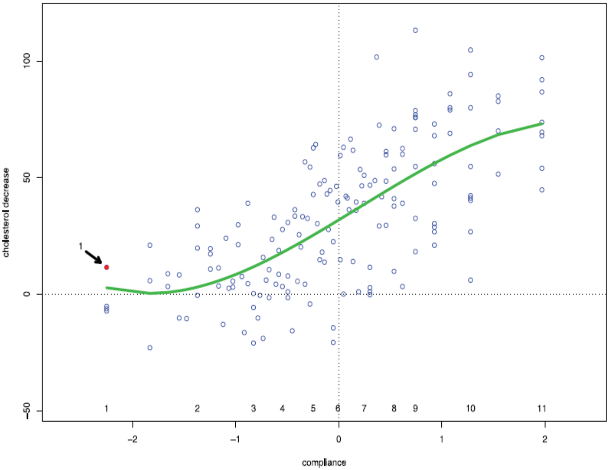


Figure 1. *Cholesterol data* Cholesterol decrease plotted versus adjusted compliance for 164 men in Treatment arm of the cholostyramine study (Efron and Feldman 1991). Solid curve is OLS cubic regression, as selected by the C_p criterion. How accurate is the curve, taking account of model selection as well as least squares fitting? (Solid arrowed point is Subject 1, featured in subsequent calculations. Bottom numbers indicate compliance for the 11 subjects in the simulation trial of 5.)

Estimation after Model Selection

- My(=Bradley Efron) usual practice:
  (a) look at data
  (b) choose model (linear, quad, cubic . . . ?)
  (c) fit estimates using chosen model
  (d) calculate standard deviation by using bootstrap
  (e) analyze as if pre-chose
- Question: Are we really happy with this result?
- My(=YC) answer: (yes..)

Nonparametric bootstrap analysis

- data $\mathbf{y} = \{(c_j, d_j), j = 1, \ldots, n = 164\} = (y_1, \ldots, y_{164})$ gave original estimate
$$\hat{\mu} = X_3 \hat{\beta}_3$$
- Bootstrap data set: $\mathbf{y}^* = (y_1^*, \ldots, y_{164}^*)$ where y_j^* drawn randomly and with replacement from data:
$$\mathrm{data} \overset{\mathrm{C_p}}{\rightarrow} \mathrm{m}^* \overset{\mathrm{OLS}}{\rightarrow} \hat{\beta}_{\mathrm{m}^*} \rightarrow \hat{\mu} = \mathrm{X}_{\mathrm{m}^*} \hat{\beta}_{\mathrm{m}^*}$$
- Bootstrap replicates $B = 4000$.
- Then empirical standard deviation of B such draws,
$$\hat{sd}_B = \left[\sum_{i=1}^{B} (\mu_i^* - \mu_\bullet^*)^2 \Big/ (B-1) \right]^{1/2},$$
where $\mu_\bullet^* = \sum_{i=1}^{B} \mu_i^* / B$.

C_p for Cholesterol Data

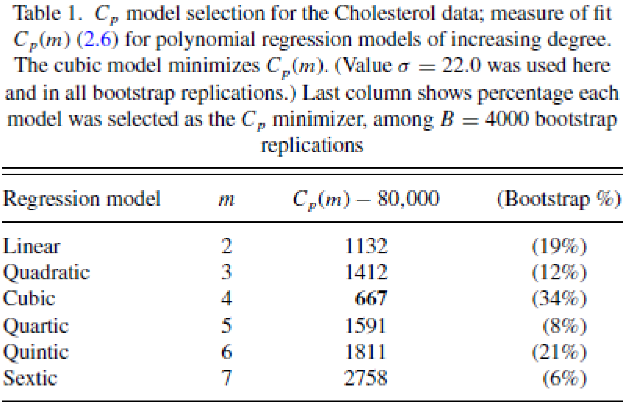


Table 1. C_p model selection for the Cholesterol data; measure of fit $C_p(m)$ (2.6) for polynomial regression models of increasing degree. The cubic model minimizes $C_p(m)$. (Value $\sigma = 22.0$ was used here and in all bootstrap replications.) Last column shows percentage each model was selected as the C_p minimizer, among $B = 4000$ bootstrap replications

Regression model	m	$C_p(m) - 80{,}000$	(Bootstrap %)
Linear	2	1132	(19%)
Quadratic	3	1412	(12%)
Cubic	4	**667**	(34%)
Quartic	5	1591	(8%)
Quintic	6	1811	(21%)
Sextic	7	2758	(6%)

Bootstrap samples

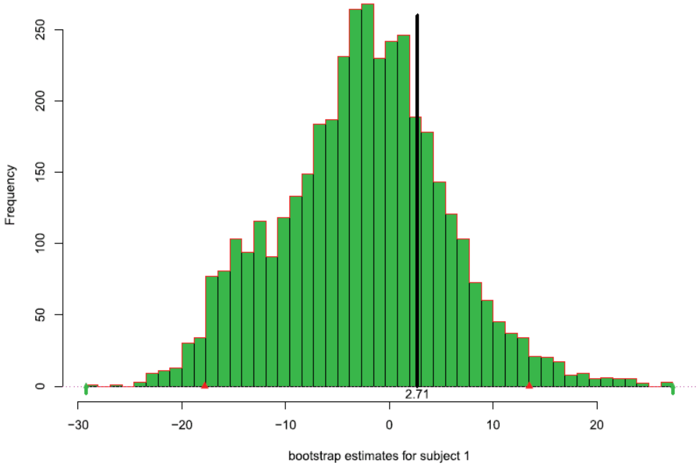


Figure 3. $B = 4000$ bootstrap replications $\hat{\mu}_1^*$ of the C_p-OLS regression estimate for Subject 1. The original estimate $t(\mathbf{y}) = \hat{\mu}_1$ is 2.71, exceeding 76% of the replications. Bootstrap standard deviation (2.4) equals 8.02. Triangles indicate 2.5th and 97.5th percentiles of the histogram.

C_p for Cholesterol Data

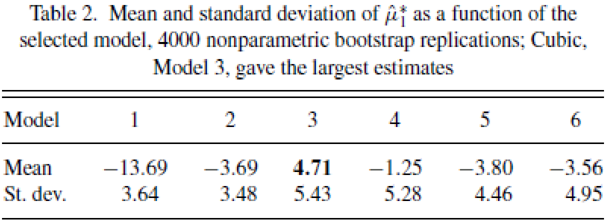


Table 2. Mean and standard deviation of $\hat{\mu}_1^*$ as a function of the selected model, 4000 nonparametric bootstrap replications; Cubic, Model 3, gave the largest estimates

Model	1	2	3	4	5	6
Mean	−13.69	−3.69	**4.71**	−1.25	−3.80	−3.56
St. dev.	3.64	3.48	5.43	5.28	4.46	4.95

- The actual dataset fell into the cubic region, giving a correspondingly large estimate: model selection can make an estimate ‘jumpy’ and erratic.

Estimation after Model Selection **AGAIN**

- My(=Bradley Efron) usual bad practice:
  (a) look at data
  (b) choose model (linear, quad, cubic . . . ?)
  (c) fit estimates using chosen model
  (d) calculate confidence interval by using bootstrap
  (e) analyze as if pre-chose
- Question1: Are we really happy with this result?
- Question2: How accurate is the fitted curve, taking account of the C_p model-selection procedure as well as OLS estimation?
- Answer: ??????????

Estimation and Accuracy after Model Selection

- Answer: Bagging (bootstrap smoothing)
- GOAL:
  Classical estimation theory ignored model selection out of necessity. Armed with modern computational equipment, statisticians can now deal with model-selection problems more realistically. The limited, but useful, goal of this article is to provide a general tool for the assessment of standard errors in such situations.

Previous researches

- Bagging (Breiman, 1996): A model-averaging device that both reduces variability and eliminates discontinuities.
- Buja and Stuetzle (2006): An excellent recent reference.
- Buhlmann and Yu (2002): Change hard thresholding estimators to soft thresholding.

Bagging

- Replace original estimator $t(\mathbf{y})$ with bootstrap average

$$\tilde{\mu} = s(\mathbf{y}) = \sum_{i=1}^{B} t(\mathbf{y}_{\mathbf{i}}^{*})/B$$

- Unlike $t(\mathbf{y})$, $s(\mathbf{y})$ does not jump as $\mathbf{y}$ crosses region boundaries, making it a more dependable vehicle for setting standard errors and confidence intervals.

Bootstrap confidence intervals

- Standard: $\hat{\mu} \pm 1.96\hat{sd}_B$
- Percentile: $[\hat{\mu}^{*(.025)}, \hat{\mu}^{*(.975)}]$
- Smoothed Standard: $\tilde{\mu} \pm 1.96\tilde{sd}_B$
  where $\tilde{sd}_B$ is a standard deviation for the smoothed bootstrap estimate $\tilde{\mu}$

Bootstrap confidence intervals

- A brute force approach employs a second level of bootstrapping: resampling y_i^* yields a collection of B second-level replications y_{ij}^{**} from which we calculate $s_i^* = \sum t(y_{ij}^{**})/B$
- Requires an enormous number of recomputations of the original statistics $t(\cdot)$.

Duplicate the paper

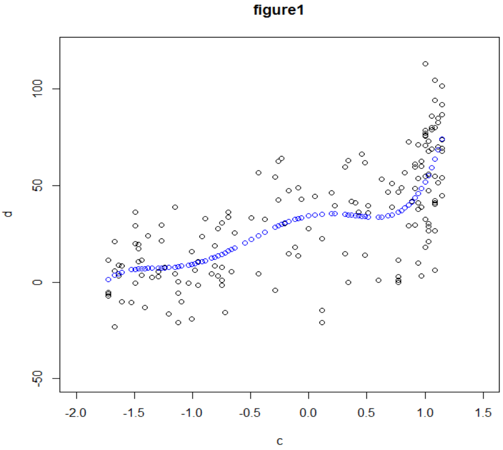


Figure: Cholesterol data, n=164 subjects: cholesterol decrease plotted versus normalized compliance; Blue points indicate OLS fifth degree polynomial

Duplicate the paper

Table: $\sigma = 22.0$ from "full model"

Model	df	$C_p - 80000$	(Boot %)
linear	2	1105	2.8
quad	3	233	3.3
cubic	4	-54	2.8
quadratic	5	-2335	15.5
quintic	**6**	**-3058**	**43.6**
sextic	7	-2744	31.6

Duplicate the paper

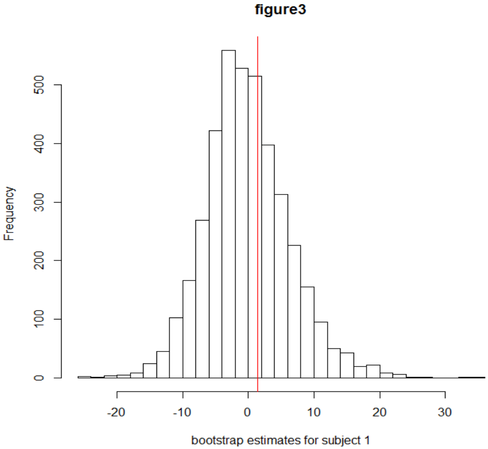


Figure: B=4000 nonparametric bootstrap replications for the model-selected regression estimate of Subject 1; 63.2% of the replications less than original

Duplicate the paper

Table: Mean and standard deviation of $\hat{\mu}_1^*$ as a function of the selected model, 4000 nonparametric bootstrap replications

Model	**1**	**2**	**3**	**4**	**5**	**6**
Mean	-1.34	5.04	-1.34	8.71	**-0.08**	-4.81
Std	2.60	2.92	3.55	5.58	**4.52**	4.78

Notations

- $s_0 = s(\mathbf{y}), t_i^* = t(\mathbf{y_i^*})$
- $Y_{ij}^* = \#\{y_{ik}^* = y_j\}$: the number of elements of $\mathbf{y_i^*}$ equaling the original data point y_j.
- The vector $Y_i^* = (Y_{i1}^*, \ldots, Y_{in}^*)$ follows a multinomial distribution with n draws on n categories each of probability $1/n$, and has mean vector and covariance matrix

$$Y_i^* \sim (\mathbf{1_n}, I_n - \mathbf{1_n}\mathbf{1_n}'/n)$$

Accuracy Theorem

Theorem

The nonparametric delta-method estimate of standard deviation for the ideal smoothed bootstrap statistic $s(\mathbf{y}) = \sum_{i=1}^{B} t(\mathbf{y_i^*})/B$ *is*

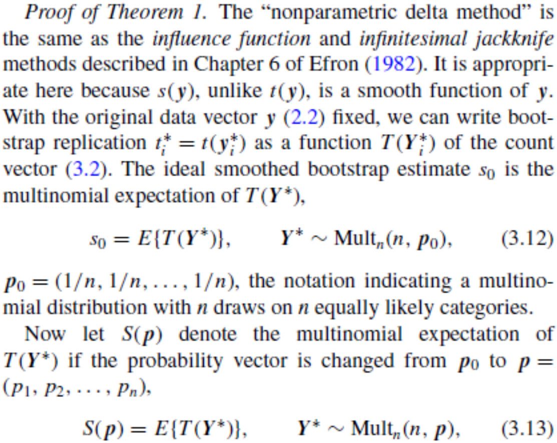
$$\tilde{sd} = \left[\sum_{j=1}^{N} cov_j^2 \right]^{1/2}$$

where

$$cov_j = cov_*(Y_{ij}^*, t_i^*),$$

the bootstrap covariance between Y_{ij}^* *and* t_i^**.*

Accuracy Theorem



Proof of Theorem 1. The "nonparametric delta method" is the same as the *influence function* and *infinitesimal jackknife* methods described in Chapter 6 of Efron (1982). It is appropriate here because $s(\boldsymbol{y})$, unlike $t(\boldsymbol{y})$, is a smooth function of $\boldsymbol{y}$. With the original data vector $\boldsymbol{y}$ (2.2) fixed, we can write bootstrap replication $t_i^* = t(\boldsymbol{y}_i^*)$ as a function $T(\boldsymbol{Y}_i^*)$ of the count vector (3.2). The ideal smoothed bootstrap estimate s_0 is the multinomial expectation of $T(\boldsymbol{Y}^*)$,

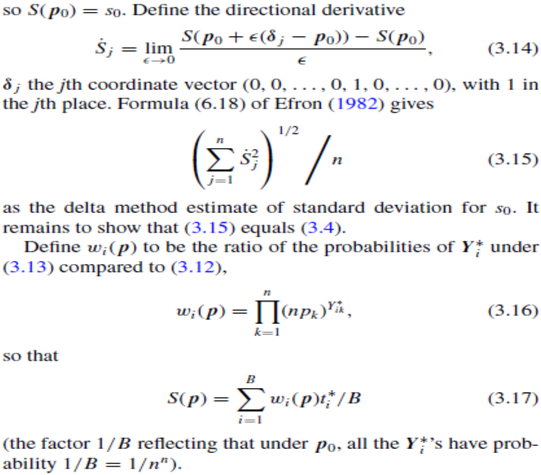
$$s_0 = E\{T(\boldsymbol{Y}^*)\}, \qquad \boldsymbol{Y}^* \sim \text{Mult}_n(n, \boldsymbol{p}_0), \qquad (3.12)$$

$\boldsymbol{p}_0 = (1/n, 1/n, \ldots, 1/n)$, the notation indicating a multinomial distribution with n draws on n equally likely categories.

Now let $S(\boldsymbol{p})$ denote the multinomial expectation of $T(\boldsymbol{Y}^*)$ if the probability vector is changed from $\boldsymbol{p}_0$ to $\boldsymbol{p} = (p_1, p_2, \ldots, p_n)$,

$$S(\boldsymbol{p}) = E\{T(\boldsymbol{Y}^*)\}, \qquad \boldsymbol{Y}^* \sim \text{Mult}_n(n, \boldsymbol{p}), \qquad (3.13)$$

Accuracy Theorem



so $S(\boldsymbol{p}_0) = s_0$. Define the directional derivative

$$\dot{S}_j = \lim_{\epsilon \to 0} \frac{S(\boldsymbol{p}_0 + \epsilon(\boldsymbol{\delta}_j - \boldsymbol{p}_0)) - S(\boldsymbol{p}_0)}{\epsilon}, \tag{3.14}$$

$\boldsymbol{\delta}_j$ the jth coordinate vector $(0, 0, \ldots, 0, 1, 0, \ldots, 0)$, with 1 in the jth place. Formula (6.18) of Efron (1982) gives

$$\left(\sum_{j=1}^{n} \dot{S}_j^2 \right)^{1/2} \Bigg/ n \tag{3.15}$$

as the delta method estimate of standard deviation for s_0. It remains to show that (3.15) equals (3.4).

Define $w_i(\boldsymbol{p})$ to be the ratio of the probabilities of $\boldsymbol{Y}_i^*$ under (3.13) compared to (3.12),

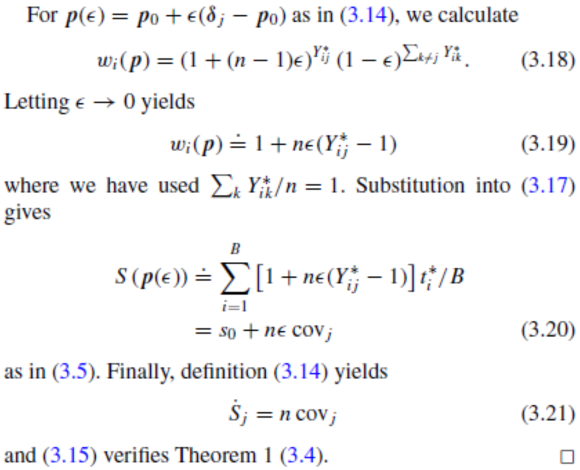
$$w_i(\boldsymbol{p}) = \prod_{k=1}^{n} (n p_k)^{Y_{ik}^*}, \tag{3.16}$$

so that

$$S(\boldsymbol{p}) = \sum_{i=1}^{B} w_i(\boldsymbol{p}) t_i^* / B \tag{3.17}$$

(the factor $1/B$ reflecting that under $\boldsymbol{p}_0$, all the $\boldsymbol{Y}_i^*$'s have probability $1/B = 1/n^n$).

Accuracy Theorem



For $p(\epsilon) = p_0 + \epsilon(\delta_j - p_0)$ as in (3.14), we calculate

$$w_i(p) = (1 + (n-1)\epsilon)^{Y_{ij}^*} (1 - \epsilon)^{\sum_{k \neq j} Y_{ik}^*}. \tag{3.18}$$

Letting $\epsilon \to 0$ yields

$$w_i(p) \doteq 1 + n\epsilon(Y_{ij}^* - 1) \tag{3.19}$$

where we have used $\sum_k Y_{ik}^*/n = 1$. Substitution into (3.17) gives

$$\begin{aligned} S(p(\epsilon)) &\doteq \sum_{i=1}^{B} \left[1 + n\epsilon(Y_{ij}^* - 1)\right] t_i^*/B \\ &= s_0 + n\epsilon \operatorname{cov}_j \end{aligned} \tag{3.20}$$

as in (3.5). Finally, definition (3.14) yields

$$\dot{S}_j = n \operatorname{cov}_j \tag{3.21}$$

and (3.15) verifies Theorem 1 (3.4). □

Sample version of Accuracy Theorem

- The estimate of standard deviation in the non-ideal case is similar:

$$\tilde{sd} = \left[\sum_{j=1}^{N} \widehat{cov}_j^2 \right]^{1/2}$$

where

$$\widehat{cov}_j = \sum_{i=1}^{B} (Y_{ij}^* - Y_{\bullet j}^*)(t_i^* - t_\bullet^*)/B$$

with $Y_{\bullet j}^* = \sum_{i=1}^{B} Y_{ij}^*/B$ and $t_\bullet^* = \sum_{i=1}^{B} t_i^*/B$

- Note that

$$\hat{sd}_B = \left[\sum_{i=1}^{B} (t_i^* - t_\bullet^*)^2 \right]^{1/2}$$

Sample version of Accuracy Theorem

- Let $\mathcal{L}(Y^*)$ be the (n-1) dimensional subspace of $\mathcal{R}^B$ spanned by the columns of the $B \times n$ matrix having elements $Y^*_{ij} - 1$.
- Also define

$$U^* = \mathbf{t} - s_o \mathbf{1}.$$

Note that U^* is the B-vector of mean-centered replications $t^*_i - s_0$.

Accuracy Theorem

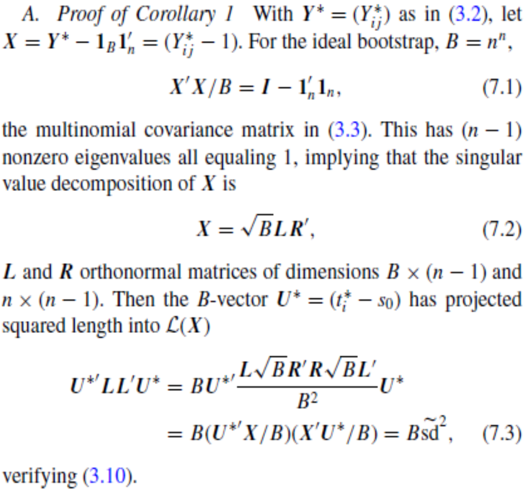
Corollary

The ratio $\tilde{sd}_B / \hat{sd}_B$ is given by

$$\frac{\tilde{sd}_B}{\hat{sd}_B} = \frac{\|\hat{U}^*\|}{\|U^*\|}$$

where $\hat{U}^*$ is the projection of U^* into $\mathcal{L}(Y^*)$.

Accuracy Theorem



A. *Proof of Corollary 1* With $\boldsymbol{Y}^* = (Y^*_{ij})$ as in (3.2), let $\boldsymbol{X} = \boldsymbol{Y}^* - \mathbf{1}_B\mathbf{1}'_n = (Y^*_{ij} - 1)$. For the ideal bootstrap, $B = n^n$,

$$\boldsymbol{X}'\boldsymbol{X}/B = \boldsymbol{I} - \mathbf{1}'_n\mathbf{1}_n, \tag{7.1}$$

the multinomial covariance matrix in (3.3). This has $(n-1)$ nonzero eigenvalues all equaling 1, implying that the singular value decomposition of $\boldsymbol{X}$ is

$$\boldsymbol{X} = \sqrt{B}\boldsymbol{L}\boldsymbol{R}', \tag{7.2}$$

$\boldsymbol{L}$ and $\boldsymbol{R}$ orthonormal matrices of dimensions $B \times (n-1)$ and $n \times (n-1)$. Then the B-vector $\boldsymbol{U}^* = (t^*_i - s_0)$ has projected squared length into $\mathcal{L}(\boldsymbol{X})$

$$\begin{aligned}\boldsymbol{U}^{*\prime}\boldsymbol{L}\boldsymbol{L}'\boldsymbol{U}^* &= B\boldsymbol{U}^{*\prime}\frac{\boldsymbol{L}\sqrt{B}\boldsymbol{R}'\boldsymbol{R}\sqrt{B}\boldsymbol{L}'}{B^2}\boldsymbol{U}^*\\ &= B(\boldsymbol{U}^{*\prime}\boldsymbol{X}/B)(\boldsymbol{X}'\boldsymbol{U}^*/B) = B\widetilde{\mathrm{sd}}^2,\end{aligned} \tag{7.3}$$

verifying (3.10).

Accuracy Theorem

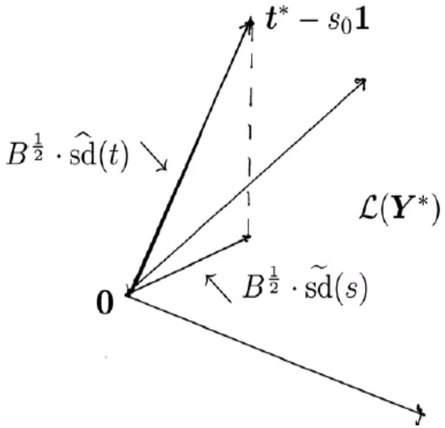

$t^* - s_0\mathbf{1}$
$B^{\frac{1}{2}} \cdot \widehat{\mathrm{sd}}(t)$ ↘
$\mathcal{L}(\mathbf{Y}^*)$
$\mathbf{0}$
↖ $B^{\frac{1}{2}} \cdot \widetilde{\mathrm{sd}}(s)$

Figure 4. Illustration of Corollary 1. The ratio $\widetilde{\mathrm{sd}}_B/\widehat{\mathrm{sd}}_B$ is the cosine of the angle between $\boldsymbol{t}^* - s_0\mathbf{1}$ (3.9) and the linear space $\mathcal{L}(\boldsymbol{Y}^*)$ spanned by the centered bootstrap counts (3.2). Model-selection estimators tend to be more nonlinear, yielding smaller ratios, that is, greater gains from smoothing.

Accuracy Theorem

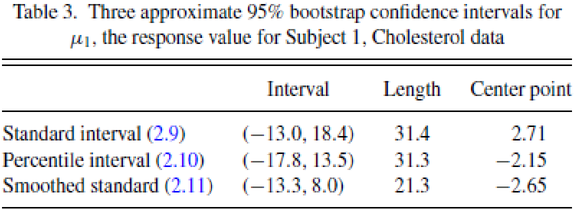


Table 3. Three approximate 95% bootstrap confidence intervals for μ_1, the response value for Subject 1, Cholesterol data

	Interval	Length	Center point
Standard interval (2.9)	(−13.0, 18.4)	31.4	2.71
Percentile interval (2.10)	(−17.8, 13.5)	31.3	−2.15
Smoothed standard (2.11)	(−13.3, 8.0)	21.3	−2.65

Accuracy Theorem

Table: Three approximate 95% bootstrap confidence intervals for μ_1, the response value for Subject 1, Cholesterol data

	Left	Right	Center point
Standard interval	-11.270996	14.135120	1.432062
Percentile interval	-11.822923	14.223378	1.200228
Smoothed standard	-9.445251	9.219275	-0.112988

Accuracy Theorem

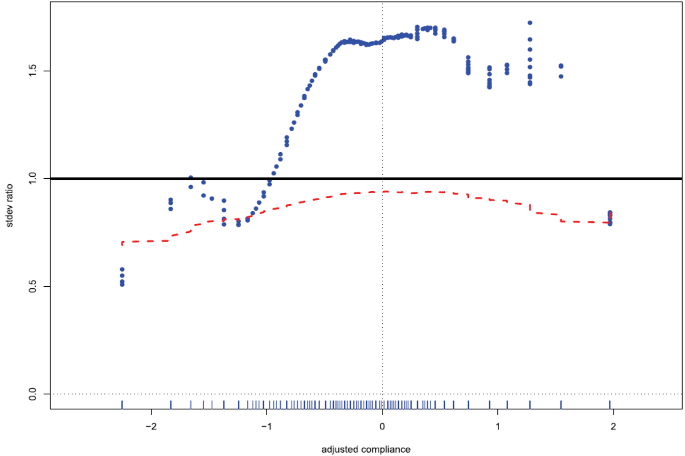


Figure 2. *Solid points*: ratio of standard deviations, taking account of model selection or not, for the 164 values $\hat{\mu}_j$ from the regression curve in Figure 1. Median ratio equals 1.52. Standard deviations including model selection are the smoothed bootstrap estimates $\widetilde{\mathrm{sd}}_B$ of Section 3. *Dashed line*: ratio of $\widetilde{\mathrm{sd}}_B$ to $\widehat{\mathrm{sd}}_B$, the unsmoothed bootstrap sd estimates as in (2.4), median 0.91.

How many bootstrap replications B are necessary?

- The jackknife provides a quick answer: divide the B replications into J groups of size B/J each, and let $\tilde{sd}_{Bj}$ be the estimate for $\tilde{sd}_B$ computed with the jth group removed.

  $$\tilde{cv}_B = \left[\frac{J}{J-1} \sum_{j=1}^{J} (\tilde{sd}_{Bj} - \tilde{sd}_{B\bullet})^2 \right]^{1/2} \Big/ \tilde{sd}_B$$

  where $\tilde{sd}_{B\bullet} = \sum \tilde{sd}_{Bj}/J$, is the jackknife estimated coefficient of variation for $\tilde{sd}_B$.
- For $B = 4000$, $\tilde{cv}_B = 0.02$ and it would have been quite sufficient.

Background

- Assume

  $$f_\alpha(\hat{\beta}) = e^{\alpha'\hat{\beta}-\psi(\alpha)}f_0(\hat{\beta})$$

  where α is the p-dimensional canonical parameter vector, $\hat{\beta}$ the p-dimensional sufficient statistic vector (playing the role of **y**), $\psi(\alpha)$ the cumulant generating function, and $f_0(\hat{\beta})$ the "carrying density".
- Under mild conditions, the expectation parameter $\beta = E_\alpha(\hat{\beta})$ is a one-to-one function of α, say $\beta = \lambda(\alpha)$, having $p \times p$ derivative matrix

  $$\frac{d\beta}{d\alpha} = V(\alpha),$$

  where V is the covariance matrix.
- Then the maximum likelihood estimate (MLE) of α is obtained by $\lambda^{-1}(\hat{\beta})$.

Parametric bootstrap

- A parametric bootstrap sample is obtained by drawing iid realizations $\hat{\beta}^*$ from the MLE density $f_{\hat{\alpha}}(\cdot)$.
- If $\hat{\mu} = t(\hat{\beta})$ is an estimate of a parameter interest μ, we have a smoothed estimate,
$$\tilde{\mu} = \sum_{i=1}^{B} t(\hat{\beta}_i^*)/B.$$
- When $t(\cdot)$ involves model selection, $\hat{\mu}$ is liable to an erratic, jumpiness, smoothed out by the averaging process.

Parametric bootstrap

- Let $\mathbb{B}$ be the $B \times p$ matrix with ith row $\hat{\beta}_i^* - \hat{\beta}$. Then

$$\mathbb{B}^{'}\mathbf{1_B}/B = \mathbb{O}, \qquad \mathbb{B}^{'}\mathbb{B} = \hat{V}.$$

- As the procedure for nonparametric bootstrap,

$$cov_* = \mathbb{B}^{'}(\mathbf{t}^* - s_0\mathbf{1_B})/B$$

Accuracy theorem for parametric version

Theorem

The parametric delta-method estimate of standard deviation for the ideal smoothed estimate is

$$\tilde{sd} = [cov_*^{`}\hat{V}^{-1}cov_*]^{1/2}$$

Corollary

For

$$\hat{sd} = [\|(\mathbf{t}^* - s_0\mathbf{1_B})\|^2/B]^{1/2},$$

the ratio $\tilde{sd}/\hat{sd}$ *is given by*

$$B^{1/2}[(\mathbf{t}^* - s_0\mathbf{1_B})^{`}\mathbb{B}(\mathbb{B}'\mathbb{B})^{-1}\mathbb{B}'(\mathbf{t}^* - s_0\mathbf{1_B})]^{1/2}/\hat{sd}$$

Schematic diagram of estimation without model selection

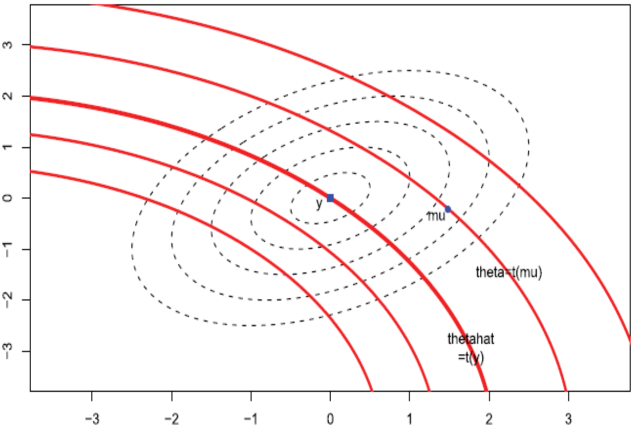


Figure 8. Schematic diagram of large-sample bootstrap estimation in situations without model selection. Observed vector $\boldsymbol{y}$ has expectation $\boldsymbol{\mu}$. Ellipses indicate bootstrap distribution of $\boldsymbol{y}^*$ given $\hat{\boldsymbol{\mu}} = \boldsymbol{y}$. Parameter of interest $\theta = t(\boldsymbol{\mu})$ is estimated by $\hat{\theta} = t(\boldsymbol{y})$. Solid curves indicate surfaces of constant value of $t(\cdot)$.

Estimation after model selection

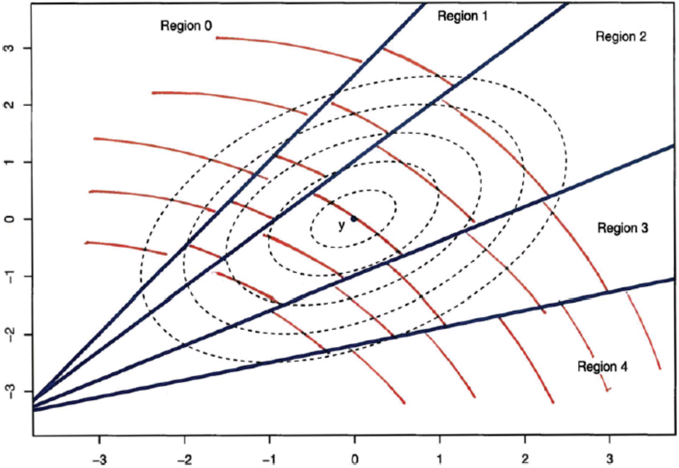


Figure 9. Estimation after model selection. The regions indicate different model choices. Now the curves of constant estimation jump discontinuously as *y* crosses regional boundaries.

What we covered

- Bootstrap!

What we didn't cover or questions

- Discrete random variable (or repeated random variable)
- Better bootstrap confidence intervals
- Asymptotic performance
- Real data problems