Model Selection, Estimation, and Bootstrap Smoothing

Bradley Efron
Stanford University
Estimation After Model Selection

• Usually:
  (a) look at data
  (b) choose model (linear, quad, cubic . . . ?)
  (c) fit estimates using chosen model
  (d) analyze as if pre-chosen

• Today: include model selection process in the analysis

• Question:
  Effects on standard errors, confidence intervals, etc.?

• Two Examples: nonparametric, parametric
Cholesterol Data

• $n = 164$ men took Cholestyramine for $\sim 7$ years

• $x = \text{compliance measure}$ (adjusted: $x \sim \mathcal{N}(0, 1)$)

• $y = \text{cholesterol decrease}$

• Regression $y$ on $x$?

  [wish to estimate: $\mu_j = E\{y|x_j\}, \ j = 1, 2, \ldots, n$]
Cholesterol data, n=164 subjects: cholesterol decrease plotted versus adjusted compliance; Green curve is OLS cubic regression; Red points indicate 5 featured subjects.
\textbf{C}_p \text{ Selection Criterion}

- \textit{Regression Model} \quad \mathbf{y} = \mathbf{X} \mathbf{\beta} + \mathbf{e} \quad \left[ e_i \sim (0, \sigma^2) \right]

- \textit{C}_p \text{ Criterion} \quad \left\| \mathbf{y} - \mathbf{X}\hat{\mathbf{\beta}} \right\|^2 + 2m\sigma^2

\hat{\beta} = \text{OLS estimate}, \quad m = \text{“degrees of freedom”}

- \textit{Model Selection:} From possible models \( X_1, X_2, X_3, \ldots \) choose the one minimizing \( C_p \).

- Then use OLS estimate from chosen model.
C<sub>p</sub> for Cholesterol Data

<table>
<thead>
<tr>
<th>Model</th>
<th>df</th>
<th>C&lt;sub&gt;p&lt;/sub&gt; − 80000</th>
</tr>
</thead>
<tbody>
<tr>
<td>M&lt;sub&gt;1&lt;/sub&gt; (linear)</td>
<td>2</td>
<td>1132</td>
</tr>
<tr>
<td>M&lt;sub&gt;2&lt;/sub&gt; (quad)</td>
<td>3</td>
<td>1412</td>
</tr>
<tr>
<td>M&lt;sub&gt;3&lt;/sub&gt; (cubic)</td>
<td>4</td>
<td>667</td>
</tr>
<tr>
<td>M&lt;sub&gt;4&lt;/sub&gt; (quartic)</td>
<td>5</td>
<td>1591</td>
</tr>
<tr>
<td>M&lt;sub&gt;5&lt;/sub&gt; (quintic)</td>
<td>6</td>
<td>1811</td>
</tr>
<tr>
<td>M&lt;sub&gt;6&lt;/sub&gt; (sextic)</td>
<td>7</td>
<td>2758</td>
</tr>
</tbody>
</table>

(σ = 22 from “full model” M<sub>6</sub>)
Nonparametric Bootstrap Analysis

• data = \{(x_i, y_i), \ i = 1, 2, \ldots, n = 164\} gave original estimate

\[ \hat{\mu} = X_3\hat{\beta}_3 \]

• Bootstrap data set \( \text{data}^* = \{(x_j, y_j)^*, \ j = 1, 2, \ldots, n\} \) where \((x_j, y_j)^*\) drawn randomly and with replacement from data:

\[
\begin{align*}
\text{data}^* & \rightarrow m^* \rightarrow \hat{\beta}_{m^*} \rightarrow \hat{\mu}^* = X_{m^*}\hat{\beta}_{m^*}
\end{align*}
\]

• I did this all \( B = 4000 \) times.
B=4000 nonparametric bootstrap replications for the model-selected regression estimate of Subject 1; boot (m,stdev)=(-2.63,8.02); 76% of the replications less than original estimate 2.71

Red triangles are 2.5th and 97.5th boot percentiles

Model Selection · Estimation · Bootstrap Smoothing
Boxplot of Cp boot estimates for Subject 1; B=4000 bootreps; Red bars indicate selection proportions for Models 1–6

only 1/3 of the bootstrap replications chose Model 3

selected model

subject 1 estimates

Model Selection · Estimation · Bootstrap Smoothing
Bootstrap Confidence Intervals

• Standard: $\hat{\mu} \pm 1.96 \hat{\text{se}}$

• Percentile: $[\hat{\mu}^{.025}, \hat{\mu}^{.975}]$

• Smoothed Standard: $\bar{\mu} \pm 1.96 \bar{\text{se}}$

• BCa/ABC: corrects percentiles for bias and changing se
95% Bootstrap Confidence Intervals for Subject 1

Confidence Interval
●
●
●
●
●
●

Standard
(−13.0,18.4)

Percentile
(−17.8,13.5)

Smoothed
(−13.3,8.0)
Bootstrap Smoothing

• **Idea**  Replace original estimator $t(y)$ with bootstrap average

\[
s(y) = \frac{1}{B} \sum_{i=1}^{B} t(y_i^*)
\]

• Same as **bagging** ("bootstrap aggregation" Breiman)

• Removes discontinuities  • Reduces variance
Accuracy Theorem

- **Notation** \( s_0 = s(y), \quad t_i^* = t(y_i^*), \ i = 1, 2, \ldots B \)
- \( Y_{ij}^* = \# \text{ of times } j\text{th data point appears in } i\text{th boot sample} \)
- \( \text{cov}_j = \sum_{i=1}^{B} Y_{ij}^* \cdot (t_i^* - s_0) / B \quad \left[ \text{covariance } Y_{ij}^* \text{ with } t_i^* \right] \)

**Theorem**  *The delta method standard deviation estimate for \( s_0 \) is*

\[
\hat{sd} = \left[ \sum_{j=1}^{n} \text{cov}_j^2 \right]^{1/2},
\]

always \( \leq \left[ \sum_{i=1}^{B} (t_i^* - s_0)^2 / B \right]^{1/2} \), the boot stdev for \( t(y) \).
Projection Interpretation

\[ \tilde{sd}(t) \rightarrow t^* - s_0 1 \]

\[ \tilde{sd}(s) \rightarrow \mathcal{L}(Y^*) \]
Standard Deviation of smoothed estimate relative to original (Red) for five subjects; green line is stdev Naive Cubic Model.

<table>
<thead>
<tr>
<th>Subject number</th>
<th>Relative stdev</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7.9</td>
</tr>
<tr>
<td>2</td>
<td>3.9</td>
</tr>
<tr>
<td>3</td>
<td>4.1</td>
</tr>
<tr>
<td>4</td>
<td>4.7</td>
</tr>
<tr>
<td>5</td>
<td>6.8</td>
</tr>
</tbody>
</table>

Bottom numbers show original standard deviations.
The Supernova Data

- **data** = \{(x_j, y_j), \ j = 1, 2, \ldots, n = 39\}
- **y_j** = absolute magnitude of Type Ia supernova
- **x_j** = vector of 10 spectral energies (350–850nm)
- **Full Model** \[ y = X_{39 \times 10} \beta + e \quad \left[ e_i^{\text{ind}} \sim \mathcal{N}(0, 1) \right] \]
Ordinary Least Squares Prediction

• Full Model \( y \sim \mathcal{N}_{39}(X\beta, I) \)

• OLS Estimates \( \hat{\mu}_{\text{OLS}} = X\hat{\beta}_{\text{OLS}} \) \( \left[ \text{arg min } \|y - X\beta\|^2 \right] \)

• Naive \( R^2 = 0.82 \) \( \left[ = \text{cor}(\hat{\mu}_{\text{OLS}}, y)^2 \right] \)

• Adjusted \( R^2 = 0.62 \) \( \left[ R^2 - \left(1 - R^2\right) \frac{m}{n - m} \right. \text{ where } m = 10 \text{ the df} \)
Adjusted absolute magnitudes for 39 Type1A supernovas plotted versus OLS predictions from 10 spectral measurements; Naive R² (squared correlation) = 0.82; Adjusted R² = 0.62.
Lasso Model Selection

- Lasso estimate is $\hat{\beta}$ minimizing $\|y - X\beta\|^2 + \lambda \sum_{1}^{p} |\beta_k|$

- Shrinks OLS estimates toward zero (all the way for some)

- Degrees of freedom "$m$" = number of nonzero $\hat{\beta}_k$'s

- Model Selection: Choose $\lambda$ (or $m$) to maximize adjusted $R^2$.

- Then $\hat{\mu} = X\hat{\beta}_m$. 
Lasso for the Supernova Data

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>$m$ (# nonzero $\hat{\beta}_k$'s)</th>
<th>Naive $R^2$</th>
<th>Adjusted $R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>63.0</td>
<td>1</td>
<td>.17</td>
<td>.12</td>
</tr>
<tr>
<td>12.9</td>
<td>4</td>
<td>.77</td>
<td>.72</td>
</tr>
<tr>
<td>3.56</td>
<td>7</td>
<td>.81</td>
<td>.73 (Selected)</td>
</tr>
<tr>
<td>0.50</td>
<td>9</td>
<td>.82</td>
<td>.71</td>
</tr>
<tr>
<td>0</td>
<td>10</td>
<td>.82</td>
<td>.62 (OLS)</td>
</tr>
</tbody>
</table>
Parametric Bootstrap Smoothing

• **Original Estimates**

\[ y \overset{\text{Lasso}}{\longrightarrow} m, \quad \hat{\beta}_m \longrightarrow \hat{\mu} = X\hat{\beta}_m \]

• **Full Model Bootstrap**

\[ y^* \sim N_{39}(\hat{\mu}_{\text{OLS}}, I) \]

\[ y^* \longrightarrow m^*, \quad \hat{\beta}_{m^*} \longrightarrow \hat{\mu}^* = X\hat{\beta}_{m^*} \]

• I did this all \( B = 4000 \) times.

• \( t^*_{ik} = \hat{\mu}^*_{ik} \)

• **Smoothed Estimates**

\[ s_k = \sum_{i=1}^{4000} t^*_{ik} / 4000 \quad [k = 1, 2, \ldots, 39] \]
Parametric Accuracy Theorem

**Theorem**  The delta method standard deviation estimate for $s_k$ is

$$\hat{sd}_k = \left[ \text{cov}_k' G \text{cov}_k \right]^{1/2},$$

where $G = X'X/B$ and $\text{cov}_k$ is bootstrap covariance between $\hat{\beta}_{OLS}$ and $t_{k}^*$. 

- Always less than the bootstrap estimate of stdev for $t_k$
- Projection into $L(\hat{\beta}_{OLS})$
Standard Deviation of smoothed estimate relative to original (Red)
for five Supernova; green line using Bootstrap reweighting

<table>
<thead>
<tr>
<th>Supernova number</th>
<th>Relative stdev</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.37</td>
</tr>
<tr>
<td>2</td>
<td>0.36</td>
</tr>
<tr>
<td>3</td>
<td>0.53</td>
</tr>
<tr>
<td>4</td>
<td>0.24</td>
</tr>
<tr>
<td>5</td>
<td>0.6</td>
</tr>
</tbody>
</table>

bottom numbers show original standard deviations
Bootstrapping the Smoothed Estimates

• Originally

\[ y \rightarrow \text{Data}^* = \{t^*_i, \quad i = 1, 2, \ldots, B\} \rightarrow s_k = \sum t^*_i / B \]

• Brute Force

\[ y^* \sim \mathcal{N}(\hat{\mu}_{\text{OLS}}, I) \rightarrow \text{Data}^{**} \rightarrow s^*_k = \sum t^{**}_i / B \]

• Requires \( B \) new bootstrap replications per “\( s^*_k \)”

• Idea: Reweight the \( B \) values in Data*
Bootstrap Reweighting

- **Smoothed Estimate** \( y \rightarrow \{ t_{ik}^* \} \rightarrow s_k = \sum t_{ik}^* / B \)
- **Reweighting** \( y^* \) gives weights \( \{ w_i^*, i = 1, \ldots, B \} \) such that
  \[
  s_k^* = \frac{\sum w_i^* t_{ik}^*}{\sum w_i^*}
  \]
- **Green Curve on Slide 22** boot standard deviation of 1000 \( s_k^* \) values [from 1000 choices of \( y^* \sim N(\hat{\mu}_{OLS}, I) \)\]
Better Confidence Intervals

• Smoothed standard intervals and percentile intervals have coverage errors of order $O\left(1/\sqrt{n}\right)$.

• “ABC” intervals have errors $O(1/n)$: corrects for bias and “acceleration” (change in stdev as estimate varies).

• Uses local reweighting for 2nd order correction
95% Confidence Intervals for 5 selected Supernovas, after subtracting original Lasso estimates; ABC (black), smoothed standard (red); green points smoothed ests
Brute Force Simulation

- Sample 500 times: \( y^* \sim N(\hat{\mu}_{\text{OLS}}, I) \); gives \( \hat{\mu}_{\text{OLS}}^* \)
- Resample \( B = 1000 \) times: \( y^{**} \sim N(\hat{\mu}_{\text{OLS}}^*, I) \)
- Use ABC to get \( \tilde{\text{sd}}, \tilde{\text{bias}}, \tilde{\text{acceleration}} \)
- Calculate ABC coverage of one-sided interval \( (-\infty, s_k) \)
  \[ s_k \text{ the original smoothed estimate} \]
- Should be uniform \([0, 1]\)
K=500 Brute force sims, B=1000 bootreps each; SN1

Supernova 2

Supernova 3

Supernova 4

Model Selection · Estimation · Bootstrap Smoothing


References


intervals. *Statist. Sci.* 11: 189–228, with comments and a rejoinder by the authors.
