On valid post-selection prediction in regression[1]

Ryan Martin
North Carolina State University
www4.stat.ncsu.edu/~rmartin

2nd Workshop on HOA & PSI
St. Louis, MO
08/13/2017

[1] Joint work with Liang Hong and Todd Kuffner.

Introduction

- Consider the usual normal linear regression model

$$y = X\beta + \sigma\varepsilon,$$

  where
  - y is $n \times 1$;
  - X is $n \times p$, with $p < n$, random design;
  - $\varepsilon \sim \mathrm{N}_n(0, I)$;
  - $\beta \in \mathbb{R}^p$ and $\sigma > 0$ are unknown.
- Even in the "low-dimensional" case, it is common to consider sub-models indexed by $S \subseteq \{1, 2, \ldots, p\}$, i.e.,

$$y = X_S\beta_S + \sigma\varepsilon.$$

- Textbooks suggest using data to choose S and carrying out inference, prediction, etc based only on the chosen S.

Introduction, cont.

- BUT, a data-driven choice of S introduces a bias that affects the classical distribution theory.
- A goal of post-selection inference to understand this *selection effect* and then to adjust for it.
- A first step is to identify the inferential target, not obvious because parameter β_S depends on S...
- I get to side-step this question because my focus is on predicting a new $\tilde{y}$ corresponding to a new $\tilde{x} \in \mathbb{R}^p$.
- *My question:* how are the standard predictive distributions affected by a data-driven choice of S?

This talk

- prediction methods
- selection rules
- (naive) bootstrap adjustment
- some simulations to see what happens:
  - AIC-based prediction intervals are not valid;
  - naive bootstrap adjustment makes problem worse.
- explanation of the phenomena
- questions and conclusions

Prediction

- For a generic S, define a "predictive distribution"

$$\tilde{y} \sim \tilde{x}_S^\top \hat{\beta}_S + \underbrace{\left[\hat{\sigma}^2\{1 + \tilde{x}_S^\top (X_S^\top X_S)^{-1} \tilde{x}_S\}\right]^{1/2}}_{\hat{\tau}_S(\tilde{x})} T,$$

  where $T \sim \mathrm{t}_{n-|S|}$.
- Summarize this via a *plausibility function*[2]

$$p_y(\tilde{y} \mid S) = 1 - \left| 2F\left(\frac{\tilde{y} - \tilde{x}_S^\top \hat{\beta}_S}{\hat{\tau}_S(\tilde{x})}\right) - 1 \right|,$$

  where F is $\mathrm{t}_{n-|S|}$ distribution function.
- $p_y(\tilde{y} \mid S)$ is not a density function, but is better in the sense that its vertical scale is meaningful.

[2] See `arXiv:1606.02352`, `arXiv:1707.00486`, etc.

Prediction, cont.

- Indeed, from the usual properties of least squares, if $S^\star$ is the oracle true model, then

$$\mathrm{P}\{p_y(\tilde{y} \mid S^\star) > \alpha\} = 1 - \alpha.$$

- This implies that upper α-level sets of the plausibility function for the oracle $S^\star$ are exact $100(1-\alpha)\%$ prediction intervals.
- Hence, prediction based on $p_y(\tilde{y} \mid S^\star)$ is *valid*.
- Same holds for the full model, though will be conservative.
- Textbooks suggest plugging in a data-driven $\hat{S}$.
- Is prediction based on $p_y(\tilde{y} \mid \hat{S})$ valid?

Selection rules: AIC

- Lots of available selection rules:
  - AIC, BIC, $*$IC, ...
  - lasso
  - ...
- Here I focus on AIC, motivated by prediction.
- In particular, $\hat{S} = \hat{S}_{\text{AIC}}$ minimizes

$$\gamma(S) = n \log\{\text{SSE}(S)\} + 2|S|,$$

  where $\text{SSE}(S)$ is the error sum of squares for model with X_S.
- Compute $\hat{S}$ via `regsubsets` in R.
- AIC is standard, but may not be "best" in any sense; certainly could do similar investigations for other selection rules.

AIC and post-selection prediction

- Straightforward to get $p_y(\tilde{y} \mid \hat{S})$ and to read off the corresponding prediction interval, etc.
- The question is if the prediction suffers somehow, e.g., the prediction intervals over- or under-cover.
- Of course, if $\hat{S} = S^\star$, then all is good.
- But AIC tends to overfit, i.e., $\hat{S} \supset S^\star$.
- Maybe overfitting isn't a problem for valid prediction.

Bootstrap adjustment

- In anticipation of problems with post-selection prediction via AIC, consider a naive bootstrap adjustment.
- Various bootstrapping approaches would probably work, here I use a variation on the method in Davison & Hinkley.
- Output is a set of samples $\{\tilde{y}^{(b)} : b = 1, \ldots, B\}$.
- Then get a predictive plausibility function, $\hat{p}_y(\tilde{y} \mid \hat{S})$, from the empirical distribution of these samples.
- Better bootstrap methods?

First experiments

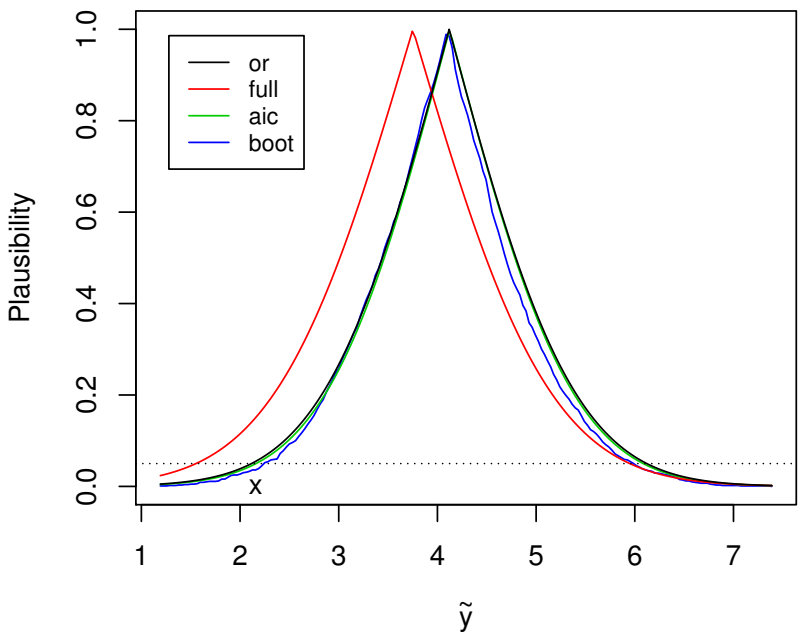
- Setup:
  - $n = 50$, $p = 10$, and $|S^\star| = 3$.
  - Rows of X are AR(1) with $\rho = 0.5$.
  - $\sigma = 1$.
- There tends to be two cases:

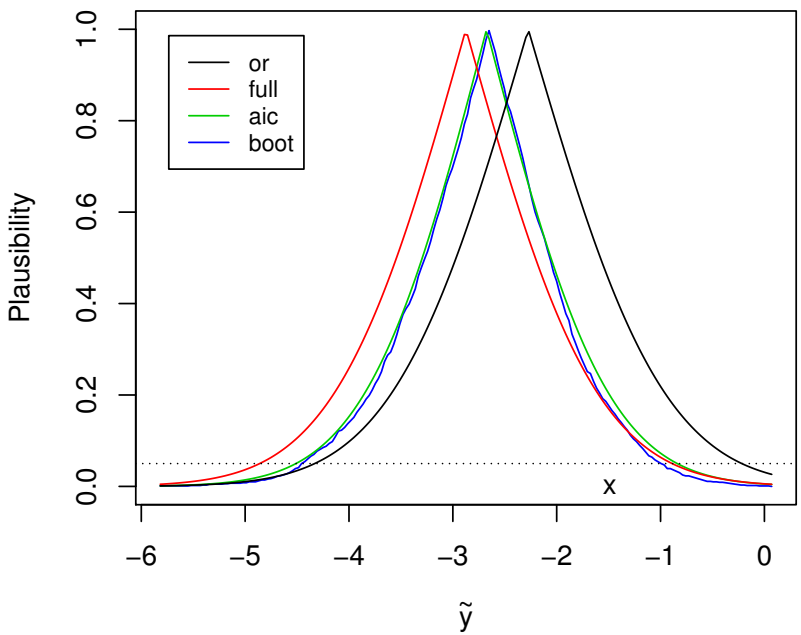
  Similar. $\hat{S} = S^\star$ so oracle and AIC plausibilities match.

  Different. $\hat{S} \neq S^\star$, and there is a location shift.

First experiments, cont.

- Examples of the similar (left) and different (right) cases.
- Notes:
  - full model is wide
  - AIC and AIC-boot look very similar



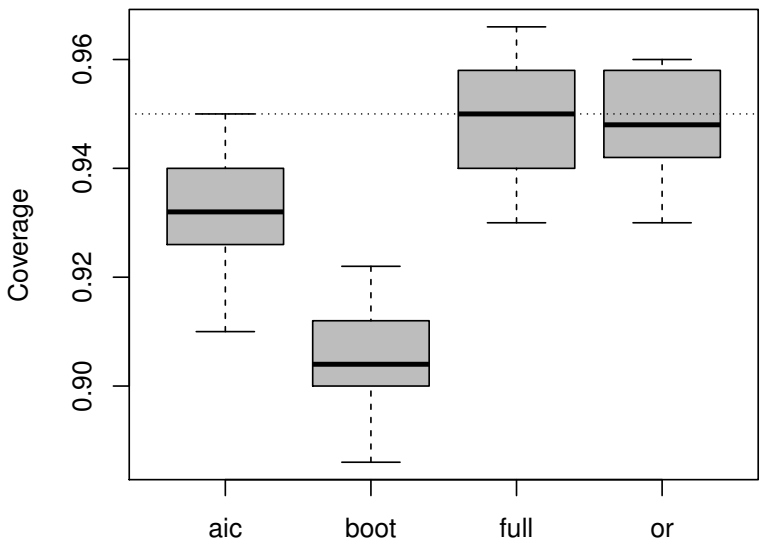


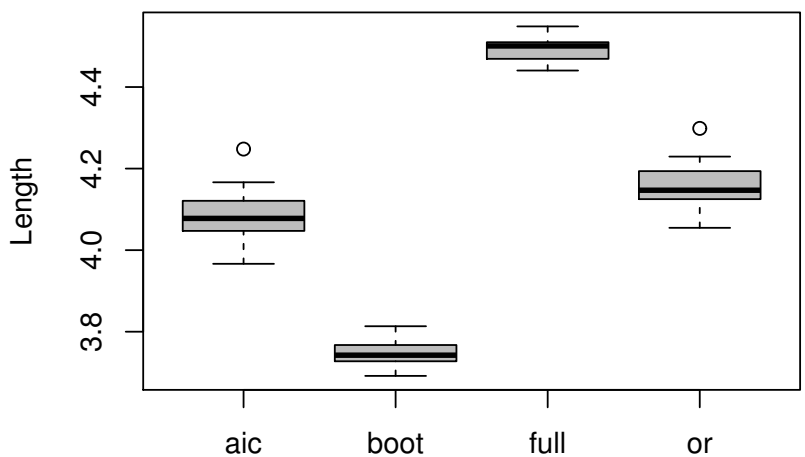
More experiments

- To see better what kind of influence selection has, let's do some more in-depth experiments.
- Look at coverage and length of prediction intervals.
- Same settings as before:
  - $n = 50$, $p = 10$, and $|S^\star| = 3$.
  - Rows of X are AR(1) with $\rho = 0.5$.
  - $\sigma = 1$.
- Repeat many times.
- Aggregate across different βs in a *moderate signal* regime; conclusions for *strong* and *weak* regimes are similar.

More experiments, cont.

- Coverage and length of 95% prediction intervals
- Notes:
  - oracle is on target
  - full model is conservative
  - AIC under-covers a bit and is slightly too short
  - AIC-boot is worse.





Questions

- It is maybe not surprising that the selection effect influences the coverage properties of the AIC based predictions.
- Counter-intuitive (to me at least) that bootstrap adjustment could somehow make things *worse*.
- Questions:
  - Any explanation for these observations?
  - Can this be fixed? If so, how?
- I can say something only about the first question.

Answers?

Proposition.

For sufficiently large n, if $\hat{S}$ is based on an AIC-like rule, then

$$\hat{S} \supset S^\star \implies \hat{\sigma}^2_{\hat{S}} < \hat{\sigma}^2_{S^\star}.$$

- AIC tends to overfit, so the proposition's condition applies.
- $\hat{\sigma}^2$ is relevant to the properties of prediction interval.
- Other aspects are relevant too, but

$$1 + \tilde{x}_S^\top (X_S^\top X_S)^{-1} \tilde{x}_S \approx 1 + n^{-1} \tilde{x}_S \Sigma_S^{-1} \tilde{x}_S \approx 1 + o(1),$$

so spread is primarily controlled by variance estimate.

Answers?, cont.

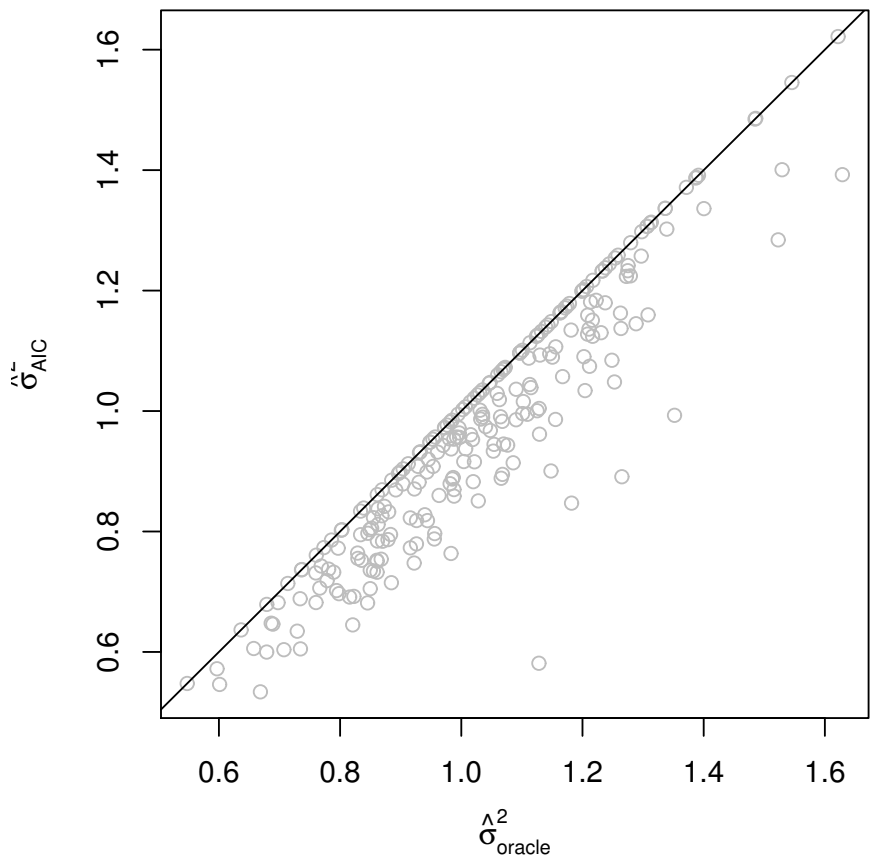
- Proof of proposition is based on SSE being involved in both AIC rule and variance estimates.
- On one side: $\hat{S} \supset S^\star$ implies

$$\mathrm{SSE}(\hat{S}) = \cdots = (1 - r_n)\mathrm{SSE}(S^\star),$$

  where r_n is connected to usual F-statistic for testing $S^\star$ vs $\hat{S}$.
- AIC selection rule puts a lower-bound on r_n.
- Plug SSE's into variance estimate formulas, and the lower-bound on r_n implies the claim.

Answers?, cont.

- Simulated 250 data sets from previous setup, and scatterplot shows the $(\hat{\sigma}^2_{S^\star}, \hat{\sigma}^2_{\hat{S}})$ pairs.



Answers?, cont.

- What about the worse performance of AIC-boot?
- Efron noted a particular "dilation property" of bootstrap.
- To paraphrase: *if an estimator is biased, then bootstrap will tend to exaggerate this*.
- He sights Neyman–Scott as an example, among others.[3]
- The above proposition indicates a bias in $\hat{\sigma}^2_{\hat{S}}$.
- Bootstrap dilation property suggests that AIC-boot intervals will be even shorter, hence worse.
- Not well understood?

[3] Can be seen in simpler examples, but results aren't remarkable...

Discussion

- Focus on a seemingly simple problem of prediction in $p < n$ normal linear regression.
- Results indicate that plugging a data-driven choice of model into the usual prediction formulas isn't valid.
- More surprisingly, a (naive) bootstrap adjustment actually makes the performance worse.
- Some theoretical explanation of these empirical findings can be given, but this is still incomplete.

Discussion, cont.

- There are certainly other methods available, including things discussed at this conference.
- My main goal was to understand the post-selection prediction problem better, and a good place to start was with the standard methods and their properties.
- It would be interesting to extend this investigation to other selection rules, e.g., lasso, and other models to see what can be said.
- These results also raise some questions about bootstrap, e.g., Efron's *dilation property*, of interest on it's own, independent of post-selection inference, etc.

The end

Thank you!