On overfitting and post-selection uncertainty assessments

Liang Hong
Department of Mathematics, Robert Morris University
hong@rmu.edu

Todd A. Kuffner
Department of Mathematics, Washington University in St. Louis
kuffner@wustl.edu

Ryan Martin
Department of Statistics, North Carolina State University
rgmarti3@ncsu.edu

Abstract

In a regression context, when the relevant subset of explanatory variables is uncertain, it is common to use a data-driven model selection procedure. Classical linear model theory, applied naively to the selected sub-model, may not be valid because it ignores the selected sub-model's dependence on the data. We provide an explanation of this phenomenon, in terms of overfitting, for a class of model selection criteria.

Keywords and phrases: Akaike information criterion; Bayesian information criterion; model selection; regression.

1 Introduction

Consider the classical multiple linear regression model

$$y = X\beta + \sigma\varepsilon, \tag{1}$$

where y is a n-vector of response variables, X is a $n \times p$ matrix of explanatory variables, β is p-vector of slope coefficients, and ε is a n-vector of independent Gaussian noise. We assume that $p < n$ and that y and the columns of X are centered so that the intercept term can be ignored. Formally, the model corresponds to the family of distributions (1) indexed by $\theta = (\beta, \sigma)$ in $\Theta = \mathbb{R}^p \times (0, \infty)$.

In practice, there is often uncertainty about the set of explanatory variables to be included. In such cases, it is common to express the parameter θ as (S, β_S, σ_S), where $S \subseteq \{1, \ldots, p\}$

represents a subset of the explanatory variables, $\beta_S \in \mathbb{R}^{|S|}$ represents the coefficients corresponding to the specific set S, and $\sigma_S > 0$. This amounts to decomposing the full parameter space Θ as $\Theta = \bigcup_S \Theta(S)$, where $\Theta(S) = \mathbb{R}^{|S|} \times (0, \infty)$. Then the model selection problem boils down to choosing a satisfactory sub-model $\Theta(S)$ or, equivalently, a subset S. Standard tools for carrying out this selection step include the Akaike information criterion, AIC (Akaike 1973), and the Bayesian information criterion, BIC (Schwarz 1978). These are designed to produce models that suitably balance parsimony and fit.

After a subset $S \subseteq \{1, \ldots, p\}$ of explanatory variables is selected, a secondary goal is to make inference on S-specific model parameters (β_S, σ_S), or functions thereof, and/or predict future values of the response. A naive approach, recommended in textbooks and commonly used by practitioners, is to replace X in (1) with X_S, the matrix with only the columns corresponding to S, and apply classical normal linear model theory. For example, for a given $x \in \mathbb{R}^p$, the classical $100(1-\alpha)\%$ confidence interval

$$C_\alpha(x; S) = x_S^\top \hat{\beta}_S \pm t_{n-|S|-1}(\alpha/2)\hat{\sigma}_S \{x_S^\top (X_S^\top X_S)^{-1} x_S\}^{1/2}, \tag{2}$$

can be used for inference on the mean response at the given x. However, as is now well-known (Berk et al. 2013), the properties that these classical procedures enjoy for a fixed/true S may not hold for a data-dependent choice, $\hat{S}$. For example, $C_\alpha(x; \hat{S})$ may not have coverage probability equal to $1 - \alpha$.

This note provides an explanation of this lack-of-validity phenomenon by showing that, when the sub-model is selected according to information criteria such as AIC and BIC, if the selected sub-model overfits, i.e., contains a superset of the explanatory variables in the true model, then the corresponding estimate of the error variance will be smaller than that for the true model. This explains the empirical findings in Hong et al. (2017), where prediction intervals based on the sub-model minimizing AIC tend to be too short compared to those based on the true model and, consequently, they tend to undercover; see Section 3. Moreover, our Theorem 1 together with the dilation phenomenon described in Efron (2003), explains why bootstrap may not correct the selection effect for methods that tend to overfit.

2 Result

For a given sub-model $\Theta(S)$, corresponding to a subset $S \subseteq \{1, \ldots, p\}$, let $(\hat{\beta}_S, \hat{\sigma}_S)$ denote the least squares estimators of the $\Theta(S)$-specific parameters (β_S, σ_S). We consider a selection procedure that chooses the subset S by minimizing the function

$$\gamma_n(S) = n \log \text{SSE}(S) + c_n |S|, \quad S \subseteq \{1, \ldots, p\}, \tag{3}$$

where $\text{SSE}(S) = \|y - X_S \hat{\beta}_S\|^2$ is the error sum of squares for sub-model $\Theta(S)$, which is proportional to the corresponding least squares estimator $\hat{\sigma}_S^2$, $c_n = o(n)$ is a user-specified sequence of constants, and $|S|$ denotes the cardinality of the set S. The AIC and BIC set $c_n \equiv 2$ and $c_n = \log n$, respectively.

Suppose that there exists a subset $S^\star$ corresponding to the truly non-zero regression coefficients, i.e., $\beta_i \neq 0$ for $i \in S^\star$ and $\beta_i = 0$ for $i \notin S^\star$. We write $(\hat{\beta}_{S^\star}, \hat{\sigma}_{S^\star})$ for the oracle estimators, those based on knowledge of the true sub-model $\Theta(S^\star)$. Of course, if $\hat{S}$ is the subset chosen by minimizing γ_n in (3), then $\gamma_n(\hat{S}) \leq \gamma_n(S^\star)$ or, equivalently,

$$n \log \text{SSE}(\hat{S}) + c_n |\hat{S}| \leq n \log \text{SSE}(S^\star) + c_n |S^\star|; \tag{4}$$

if $\hat{S} \neq S^\star$, then the inequality in (4) would be strict.

For the purpose of inference or prediction, it is common to naively use the classical normal linear model theory, based on the selected subset $\hat{S}$, to derive uncertainty assessments. However, using the data to select $\hat{S}$ introduces bias, violating the assumptions of that classical theory, and thereby invalidating the conclusions. The next result provides an explanation for this general phenomenon in cases where the selected sub-model $\Theta(\hat{S})$ overfits in the sense that $\hat{S} \supset S^\star$. In such cases, we find that $\hat{\sigma}_{\hat{S}}$ is smaller than the oracle estimator $\hat{\sigma}_{S^\star}$. Since the error variance estimate is involved in all uncertainty assessment calculations, and since it is common for selection methods to overfit, especially those based on AIC (Hurvich & Tsai 1989), this systematic under-estimation explains the general lack of validity of the classical inferential tools applied naively in a post-selection context.

Theorem 1. *Suppose* $\hat{S} \supset S^\star$. *If*

$$1 - \exp(-a_n D_n) > D_n, \tag{5}$$

where $a_n = (c_n/n)(n - |S^\star| - 1)$ *and* $D_n = (|\hat{S}| - |S^\star|)/(n - |S^\star| - 1)$, *then* $\hat{\sigma}_{\hat{S}} < \hat{\sigma}_{S^\star}$.

To gain some intuition about the condition (5), first note that $a_n D_n$ will tend to be small. In particular, a very conservative bound is $a_n D_n \leq c_n p/n$, which is small for moderate c_n and $n \gg p$. Next, since $x \mapsto 1 - \exp(-ax)$ is convex for $x > 0$ and $a > 0$, we have $1 - \exp(-a_n D_n) > a_n D_n$ for all D_n in an interval $(0, d)$, where $d = d(a_n) \in [0, 1)$. So, to meet (5) we need $a_n > 1$ and, again, we have a conservative bound $a_n \geq c_n(n - p - 1)/n$, which itself is greater than 1 for $n \gg p$ and c_n not too small. In particular, if $n \gg p$ and $c_n \equiv 2$ as in the AIC, then (5) holds.

of Theorem 1. Start by writing $\text{SSE}(\hat{S})$ in terms of $\text{SSE}(S^\star)$. Let $X_{\hat{S}}$ and $X_{S^\star}$ denote the submatrices corresponding to the indicated subsets, and write $P_{\hat{S}}$ and $P_{S^\star}$ for the respective projections onto their column spaces. Then Pythagoras' theorem implies that

$$\text{SSE}(\hat{S}) = \text{SSE}(S^\star) + Y^\top (P_{S^\star} - P_{\hat{S}}) Y = (1 - r_n)\text{SSE}(S^\star),$$

where

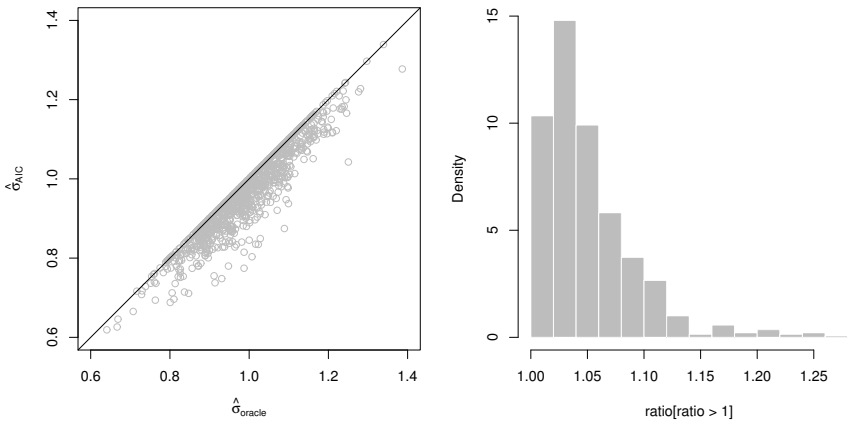
$$r_n = r_n(S^\star, \hat{S}) = \frac{|\hat{S}| - |S^\star|}{n - |\hat{S}|} F_n(S^\star, \hat{S}),$$

and $F_n(S^\star, \hat{S})$ is the usual F-statistic for testing the larger $\Theta(\hat{S})$ against the smaller $\Theta(S^\star)$. Consequently, we choose $\hat{S}$ over the strictly smaller $S^\star$, according to (4), if and only if $r_n > 1 - \exp(-a_n D_n)$.

Then the above connection between $\text{SSE}(\hat{S})$ and $\text{SSE}(S^\star)$ immediately gives a comparison between the corresponding variance estimates:

$$\hat{\sigma}^2_{\hat{S}} = \frac{\text{SSE}(\hat{S})}{n - |\hat{S}| - 1} = \frac{(1 - r_n)\text{SSE}(S^\star)}{n - |\hat{S}| - 1} = \frac{n - |S^\star| - 1}{n - |\hat{S}| - 1}(1 - r_n)\hat{\sigma}^2_{S^\star}.$$

As above, we find that $\hat{\sigma}_{\hat{S}} < \hat{\sigma}_{S^\star}$ if and only if $r_n > D_n$. By condition (5), it follows that the lower bound on r_n derived from over-fitting is greater than that derived from the under-estimation. Therefore, over-fitting implies under-estimation, proving the claim. □



(a) Scatterplot of $\hat{\sigma}_{\hat{S}}$ versus $\hat{\sigma}_{S^\star}$ (b) Histogram of $\hat{\sigma}_{S^\star}/\hat{\sigma}_{\hat{S}}$, over-fit only

Figure 1: Plots from the simulations described in Section 3.

3 Illustration

Consider the model (1), with $n = 50$ and $p = 10$, and variance $\sigma^2 = 1$. Set $S^\star = \{1, 2, 3\}$, with corresponding coefficients $\beta_1^\star = 1$, $\beta_2^\star = 2$, and $\beta_3^\star = 3$. The rows of the X matrix are independent, p-variate normal, with mean zero, AR(1) dependence structure, and one-step correlation $\rho = 0.5$. We simulated 1000 data sets and, for each, evaluated $\hat{\sigma}_{\hat{S}}$ and $\hat{\sigma}_{S^\star}$, where $\hat{S}$ is chosen based on the AIC. The scatterplot shown in Figure 1(a) demonstrates the systematic under-estimation based on the AIC-selected sub-model, as predicted by Theorem 1. In all 1000 cases, we have $\hat{S} \supseteq S^\star$, and those on the diagonal line correspond to $\hat{S} = S^\star$. To further illustrate the difference between the estimates, Figure 1(b) plots a histogram of the ratio $\hat{\sigma}_{S^\star}/\hat{\sigma}_{\hat{S}}$, only for the strict over-fit cases. In particular, the mean from this histogram is 1.06.

While the relative difference between the two estimates does not seem remarkable, even this small of a difference can impact the quality of inference. For example, consider using the confidence interval (2) for inference on the mean response at a particular setting x of the explanatory variables; here, x is an independent sample from the distribution that generated the rows of X. The oracle 95% confidence interval $C_{0.05}(x; S^\star)$ has coverage exactly equal to 0.95 but, in the 1000 simulations above, the coverage probability of $C_\alpha(x; \hat{S})$ is roughly 0.86. It happens that the $\hat{S}$-based intervals tend to be shorter than the oracle, suggesting that valid post-selection inference on the mean response requires $\hat{\sigma}_{\hat{S}}$ to be strictly larger than $\hat{\sigma}_{S^\star}$, which is impossible given Theorem 1 and the AIC's tendency to over-fit.

Acknowledgement

The authors are grateful to the Editor and two referees whose comments greatly enhanced the clarity of our presentation. Kuffner was supported by the National Science Foundation, U.S.A.

References

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, B. Petrov & F. Csáki, eds. Akadémiai Kiadó.

BERK, R., BROWN, L., BUJA, A., ZHANG, K. & ZHAO, L. (2013). Valid post-selection inference. *Ann. Statist.* **41**, 802–837.

EFRON, B. (2003). Second thoughts on the bootstrap. *Statist. Sci.* **18**, 135–140.

HONG, L., KUFFNER, T. A. & MARTIN, R. G. (2017). On prediction of future insurance claims when the model is uncertain. *Submitted*. Available at SSRN: https://ssrn.com/abstract=2883574.

HURVICH, C. M. & TSAI, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika* **76**, 297–307.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–464.