AIC post-selection inference in linear regression

Ali Charkhi*[1] and Gerda Claeskens[1]

[1] *ORSTAT and Leuven Statistics Research Center, KU Leuven, Faculty of Economics and Business, Naamsestraat 69, 3000 Leuven, Belgium*

Post-selection inference has been considered a crucial topic in data analysis. In this article, we develop a new method to obtain correct inference after model selection by the Akaike's information criterion [1] in linear regression models. Confidence intervals can be calculated by incorporating the randomness of the model selection in the distribution of the parameter estimators which act as pivotal quantities. Simulation results show the accuracy of the proposed method.

Keywords: Post-selection inference; Confidence intervals; Akaike information criterion.

1 Introduction

Consider the linear regression setting where the true model is of the form

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\epsilon} \tag{1}$$

where $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$ and we assume that σ^2 is known. For a given predictor matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p) \in \mathbb{R}^{n \times p}$, we wish to model $\boldsymbol{\mu}$ by a linear function of all predictors, $\boldsymbol{Xb}$, or just a subset of predictors, $\boldsymbol{X}_M \boldsymbol{b}_M$, where $\boldsymbol{X}_M$ contains as columnsthe predictors with indices in $M \subseteq \{1, \ldots, p\}$. This setting can be considered as a nonparametric setting because there is no assumption about whether the true model is also linear for a true coefficients vector $\boldsymbol{\beta}^0$. The least squares estimator in linear regression is defined as $\widehat{\boldsymbol{\beta}}_M = (\boldsymbol{X}_M^t \boldsymbol{X}_M)^{-1} \boldsymbol{X}_M^t \boldsymbol{Y}$ which minimizes the expected squared error. In other words, $\widehat{\boldsymbol{\beta}}_M$ is the estimator of $\boldsymbol{\beta}_M = (\boldsymbol{X}_M^t \boldsymbol{X}_M)^{-1} \boldsymbol{X}_M^t \boldsymbol{\mu}$.

Regarding the inference, one can easily use classical confidence intervals (in any submodel) based on the normality of the observations. The difficulty arises when one selects a model based on a criterion from a collection of potential

*Corresponding author: ali.charkhi@kuleuven.be

models $\mathcal{M}$ and wants to do inference for the parameters in the selected model. Since this selection is data-driven, it is random. Ignoring this randomness may lead to incorrect inference. One way to incorporate the selection randomness in inference is using conditional inference, by conditioning on the selected model.

When one imposes the assumption that there exist a true model with parameters $\boldsymbol{\beta}^0$, [7, 8] showed that the distribution of a post-selection estimator can not be estimated in a uniform way. Considering model (1), [2] proposed a method to calculate confidence intervals which are valid irrespective of the selection criterion (Posi method), hence their confidence intervals are conservative for a specific model selection criterion. Their confidence intervals are for parameters in the selected model rather than the true value of the parameters. [6] studied post-selection inference for lasso in high dimensional data. [9] generalized the results to sequential regression procedures such as forward stepwise regression and least angle regression. [3] used the asymptotic distribution to calculate confidence intervals for the model parameters in general likelihood models when they assumed that there exits a true model (Asymp-AIC method).

In this article, we study post-selection inference for the population parameters after using AIC for model selection without assuming a true model to exist. Assuming σ^2 is known, AIC for model M is defined as

$$AIC(M) = \|\boldsymbol{Y} - \boldsymbol{X}_M\widehat{\boldsymbol{\beta}}_M\|^2 + 2\sigma^2|M|. \tag{2}$$

Knowledge about σ^2 may seem restrictive, but [5] showed that in this setting inference without knowing σ^2 is impossible. The main reason is that taking the variance estimation into account leads to insufficient information about the parameters for inference. Our simulations show that even we estimate the σ^2 using the same data, the results are still valid. When σ^2 is unknown, the AIC score for each model is different from the score in (2). In that case, one estimates σ^2 within each model by $\widehat{\sigma}^2 = \|\boldsymbol{Y} - \boldsymbol{X}_M\widehat{\boldsymbol{\beta}}_M\|^2/n$ which leads to the following formula for AIC for model M:

$$AIC(M, \sigma^2) = \log(\|\boldsymbol{Y} - \boldsymbol{X}_M\widehat{\boldsymbol{\beta}}_M\|^2) + \frac{2(|M|+1)}{n}. \tag{3}$$

In a set of models $\mathcal{M}$ the model with the smallest value of (2), or (3), is the best model according to AIC in the considered case.

2 Post-selection inference

When AIC selects a model, it defines an event which we call the *selection event.* If AIC selects model M, i.e. $M_{\text{aic}} = M$, then

$$AIC(M) \leq AIC(M_i), \qquad \forall M_i \in \mathcal{M}.$$

Define $\boldsymbol{P}_M = \boldsymbol{X}_M(\boldsymbol{X}_M^t\boldsymbol{X}_M)^{-1}\boldsymbol{X}_M^t$. Using (2) we can represent the selection event as

$$\begin{aligned} \mathcal{S}_M(\mathcal{M}) &= \bigcap_{M_i \in \mathcal{M}} \left\{ \|(\boldsymbol{I}_n - \boldsymbol{P}_{M_i})\boldsymbol{Y}\|^2 + 2\sigma^2|M_i| - \|(\boldsymbol{I}_n - \boldsymbol{P}_M)\boldsymbol{Y}\|^2 - 2\sigma^2|M| \geq 0 \right\} \\ &= \bigcap_{M_i \in \mathcal{M}} \left\{ \boldsymbol{Y}^t(\boldsymbol{P}_M - \boldsymbol{P}_{M_i})\boldsymbol{Y} - 2\sigma^2(|M| - |M_i|) \geq 0 \right\}. \qquad (4) \end{aligned}$$

Similarly when (3) is used for selection, the event can be expressed as

$$\begin{aligned} \mathcal{S}_M^{\sigma^2}(\mathcal{M}) &= \bigcap_{M_i \in \mathcal{M}} \left\{ \log\left(\frac{\left\| \boldsymbol{Y} - \boldsymbol{X}_{M_i}\widehat{\boldsymbol{\beta}}_{M_i} \right\|^2}{\left\| \boldsymbol{Y} - \boldsymbol{X}_M\widehat{\boldsymbol{\beta}}_M \right\|^2} \right) \geq \frac{2(|M| - |M_i|)}{n} \right\} \\ &= \bigcap_{M_i \in \mathcal{M}} \left\{ \boldsymbol{Y}^t(\boldsymbol{I}_n - \boldsymbol{P}_{M_i})\boldsymbol{Y} \cdot \kappa_{M_i} - \boldsymbol{Y}^t(\boldsymbol{I}_n - \boldsymbol{P}_M)\boldsymbol{Y} \cdot \kappa_M \geq 0 \right\}, (5) \end{aligned}$$

where $\kappa_{M_i} = \exp\left(2(|M_i|)/n\right)$.

To obtain correct confidence intervals after model selection, we use conditional inference. In other words, for inference for a parameter of the form $\boldsymbol{\eta}_M^t\boldsymbol{\mu}$ in the selected model where $\boldsymbol{\eta}_M \in \mathbb{R}^n$ and is specified, we need to investigate the distribution of $\boldsymbol{\eta}_M^t\boldsymbol{Y} \mid \{M_{\text{aic}} = M\}$ which is equivalent to working with

$$\boldsymbol{\eta}_M^t\boldsymbol{Y} \mid \mathcal{S}_M(\mathcal{M}).$$

It is possible to rewrite $\mathcal{S}_M(\mathcal{M})$ in terms of $\boldsymbol{\eta}_M^t\boldsymbol{Y}$. Proofs for the following results can be found in [4].

Lemma 1. *Define $T = \boldsymbol{\eta}_M^t\boldsymbol{Y}$ and $\boldsymbol{Z} = \boldsymbol{Y} - \boldsymbol{w}T$ where $\boldsymbol{w} = \boldsymbol{\eta}_M(\boldsymbol{\eta}_M^t\boldsymbol{\eta}_M)^{-1}$ (T and $\boldsymbol{Z}$ are independent). Then*

$$\begin{aligned} \mathcal{S}_M(\mathcal{M}) = \bigcap_{M_i \in \mathcal{M}} \{ \qquad & T^t\boldsymbol{w}^t\boldsymbol{D}_{M_i}\boldsymbol{w}T + 2T^t\boldsymbol{w}\boldsymbol{D}_{M_i}\boldsymbol{Z} \\ & +\boldsymbol{Z}^t\boldsymbol{D}_{M_i}\boldsymbol{Z} - 2\sigma^2(|M| - |M_i|) \geq 0\} \qquad (6) \end{aligned}$$

and

$$\begin{aligned}\mathcal{S}_M^{\sigma^2}(\mathcal{M}) = \bigcap_{M_i\in\mathcal{M}} \{ \quad & T^t\boldsymbol{w}^t\boldsymbol{R}_{M_i}\boldsymbol{w}T\kappa_{M_i} - T^t\boldsymbol{w}^t\boldsymbol{R}_M\boldsymbol{w}T\kappa_M \\ & +2T^t\boldsymbol{w}\boldsymbol{R}_{M_i}\boldsymbol{Z}\kappa_{M_i} - 2T^t\boldsymbol{w}\boldsymbol{R}_M\boldsymbol{Z}\kappa_M \\ & +\boldsymbol{Z}^t\boldsymbol{R}_{M_i}\boldsymbol{Z}\kappa_{M_i} - \boldsymbol{Z}^t\boldsymbol{R}_M\boldsymbol{Z}\kappa_M \geq 0\} \end{aligned} \qquad (7)$$

where $\boldsymbol{D}_{M_i} = \boldsymbol{P}_M - \boldsymbol{P}_{M_i}$ *and* $\boldsymbol{R}_{M_i} = \boldsymbol{I}_n - \boldsymbol{P}_{M_i}$.

As expressions (6) and (7) show, the selection event can be written via quadratic functions of T. For the selection event in (6), define

$$a_i = \boldsymbol{w}^t\boldsymbol{D}_{M_i}\boldsymbol{w}, \quad b_i = 2\boldsymbol{w}\boldsymbol{D}_{M_i}\boldsymbol{Z}, \quad c_i = \boldsymbol{Z}^t\boldsymbol{D}_{M_i}\boldsymbol{Z} - 2\sigma^2(|M| - |M_i|),$$

and for the selection event in (7),

$$a_i = \boldsymbol{w}^t\boldsymbol{R}_{M_i}\boldsymbol{w}\kappa_{M_i} - \boldsymbol{w}^t\boldsymbol{R}_M\boldsymbol{w}\kappa_M, \quad b_i = 2(\boldsymbol{w}\boldsymbol{R}_{M_i}\boldsymbol{Z}\kappa_{M_i} - \boldsymbol{w}\boldsymbol{R}_M\boldsymbol{Z}\kappa_M),$$
$$c_i = \boldsymbol{Z}^t\boldsymbol{R}_{M_i}\boldsymbol{Z}\kappa_{M_i} - \boldsymbol{Z}^t\boldsymbol{R}_M\boldsymbol{Z}\kappa_M.$$

For both selection events in (6) and (7), it is obvious that the selection event can be written as

$$\bigcap_{M_i\in\mathcal{M}} \{a_iT^2 + b_iT + c_i \geq 0\}.$$

These inequalities lead to allowable values for T, namely, of the form $I_M^{\boldsymbol{Z}}(\mathcal{M}) = \cup_{i=1}^{l}(a_i(\boldsymbol{Z}), b_i(\boldsymbol{Z}))$. So, the estimator T for the population parameter $\boldsymbol{\eta}^t\boldsymbol{\mu}$ is a normal random variable which is restricted in $I_M^{\boldsymbol{Z}}(\mathcal{M})$.

Denote the standard normal CDF by $\Phi(x)$ and also denote the CDF of a $N(\mu,\sigma^2)$ random variable truncated to $D = \cup_{i=1}^{l}(a_i, b_i)$ by $F(\cdot;\mu,\sigma^2,D)$ which can be written as, for $x \in (a_r, b_r)$

$$F(x;\mu,\sigma^2,D) = \frac{\sum_{i=1}^{r-1} p_i + \Phi((x-\mu)/\sigma) - \Phi((a_r-\mu)/\sigma)}{\sum_{i=1}^{l} p_i}, \qquad (8)$$

where $p_i = \Phi((b_i-\mu)/\sigma) - \Phi((a_i-\mu)/\sigma)$. The following result shows how we can use (8) as a pivotal quantity.

Result 1: *Let* $\boldsymbol{\eta} \in \mathbb{R}^n$ *be fixed,* $T = \boldsymbol{\eta}^t\boldsymbol{Y}$ *and the selection event is* $\mathcal{S}_M(\mathcal{M})$*, Then*

$$F\left(T;\boldsymbol{\eta}^t\boldsymbol{\mu}, \sigma^2\left\|\boldsymbol{\eta}\right\|^2, I_M^{\boldsymbol{Z}}(\mathcal{M})\right) \mid \mathcal{S}_M(\mathcal{M}) \sim \text{ Unif}\,(0,1). \qquad (9)$$

In post-selection inference, we are interested in constructing confidence intervals for parameters in the selected model. We mainly focus on a one-dimensional parameter. For parameters in the selected model, we construct confidence intervals for each parameter separately. In general, for $\boldsymbol{\eta}^t\boldsymbol{\mu} \in \mathbb{R}$ we are interested in obtaining a confidence interval $[L, U]$ such that $P(L \leq \boldsymbol{\eta}^t\boldsymbol{\mu} \leq U | I_M^{\boldsymbol{Z}}(\mathcal{M})) = 1 - \alpha$. We can use (9) to construct confidence intervals based on the method of pivotal quantities.

Result 2 *Let* $\boldsymbol{\eta} \in \mathbb{R}^n$ *and* $T = \boldsymbol{\eta}^t\boldsymbol{Y}$. *Define* L *and* U *such that*

$$F(T; L, \sigma^2 \left\|\boldsymbol{\eta}\right\|^2, I_M^{\boldsymbol{Z}}(\mathcal{M})) = 1 - \frac{\alpha}{2}, \qquad F(T; U, \sigma^2 \left\|\boldsymbol{\eta}\right\|^2, I_M^{\boldsymbol{Z}}(\mathcal{M})) = \frac{\alpha}{2},$$

then $[L, U]$ *is a confidence interval for* $\boldsymbol{\eta}^t\boldsymbol{\mu}$ *conditional on* $M_{aic} = M$ *such that* $P(\boldsymbol{\eta}^t\boldsymbol{\mu} \in [L, U] \mid M_{aic} = M) = 1 - \alpha$.

Result 2 is a general result, because $\boldsymbol{\eta} \in \mathbb{R}^n$ can be defined by the user. For instance, considering $\boldsymbol{\eta}^t = \boldsymbol{e}_i(\boldsymbol{X}_M^t\boldsymbol{X}_M)^{-1}\boldsymbol{X}_M^t$ as the direction of interest for inference, Result 2 provides a confidence interval for the ith parameter in the selected model.

If the true model is indeed linear, i.e. there exist a $\boldsymbol{\beta}^0$ such that $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}^0$, and AIC selects a model M which does not contain all non-zero parameters, then $\widehat{\boldsymbol{\beta}}$ is an unbiased estimator not for the true parameters but for

$$\boldsymbol{\beta}_M = \boldsymbol{\beta}^0[M] + (\boldsymbol{X}_M^t\boldsymbol{X}_M)^{-1}\boldsymbol{X}_M^t\boldsymbol{X}_{M^c}\boldsymbol{\beta}^0[M^c] \tag{10}$$

where M^c denotes the parameters not in the model M and $\boldsymbol{\beta}^0[M]$ represents the true coefficients in the model M. Result 2 can be used to calculate the confidence intervals for the components of $\boldsymbol{\beta}_M$.

3 Simulation study

Consider

$$Y_i = \sin(2x_i) + \epsilon_i, \qquad i = 1, \ldots, n,$$

where $x_i \sim N(0, 4)$ and $\epsilon_i \overset{i.i.d}{\sim} N(0, 9)$ for $i = 1, \ldots, 50$. In the models, consider orthogonal polynomials of degree 8. We include the intercept and the first order of the polynomial in all models and we fit all possible models with the other 7 terms (2^7 models). Denote the orthogonal polynomials by $\boldsymbol{g}(x) = (g_1(x), \ldots, g_8(x))$, we want to approximate $\sin(2x)$ by orthogonal polynomials. We run the simulation until the model with $M = (\beta_0, \beta_1, \beta_3, \beta_5)$ has been selected 1000 times. Denote $\boldsymbol{g_1} = (1_n, \boldsymbol{g})$, including a unit column for the intercept. The confidence intervals are calculated for the components of $(\boldsymbol{g_{1M}}^t\boldsymbol{g_{1M}})^{-1}\boldsymbol{g_{1M}}^t\sin(2\boldsymbol{x})$.

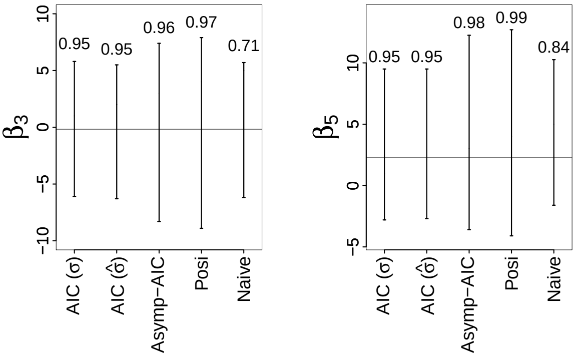

Figure 1: Mean of confidence intervals and their coverage probabilities over 1000 replications for different methods.

Figure 1 shows the mean of the confidence intervals over 1000 simulation runs for different methods along with their coverage probabilities for β_3 and β_5. We denote the proposed method by AIC(σ) when we use the knowledge about the σ and denote by AIC($\widehat{\sigma}$) where we estimate the variance in the full model. The results for [3] (Asymp-AIC) and [2] (Posi) are also presented. Both AIC(σ) and AIC($\widehat{\sigma}$) outperform other methods in terms of confidence interval lengths. The naive method leads to confidence intervals with similar length but the coverage probability is lower than the nominal value.

4 Conclusion

We proposed a new method for considering the selection randomness in inference by AIC for linear regression. In contrast the Asymp-AIC proposed by [3] which holds asymptotically, we do not need to simulate from the constrained multivariate normal distribution and the results are exact even in small sample sizes. The method performs better than PostAIC when the linear model is not the correct model. For normal linear regression models this method can be considered as a complement for PostAIC. Because if we assume the selected model is correct, the PostAIC can generate accurate confidence intervals; otherwise, the proposed method in this chapter can be used.

Acknowledgements: Your acknowledgements.

References

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. *2nd International Symposium on Information Theory*, 267–281, 1973.

[2] R. Berk, L. Brown, A. Buja, K. Zhang and L. Zhao. Valid post-selection inference. *The Annals of Statistics*, 41:802–837, 2013.

[3] A. Charkhi and G. Claeskens. Asymptotic post-selection inference for Akaike's information criterion. *Technical report.*

[4] A. Charkhi and G. Claeskens. Exact post-selection inference for AIC in linear regression. *Technical report.*

[5] W. Fithian, D. L. Sun and J. Taylor. Optimal inference after model selection. *Technical report.*

[6] J. D. Lee, D. L. Sun, Y. Sun and J. Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

[7] H. Leeb and B. M. Pötscher. Model selection and inference: Facts and fiction. *Econometric Theory*, 21:22–59, 2005.

[8] H. Leeb and B. M. Pötscher. Can one estimate the conditional distribution of post-model-selection estimators? *The Annals of Statistics*, 34:2554–2591, 2006.

[9] J. Taylor, R. Lockhart, R. J. Tibshirani, R. Tibshirani. Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111:600–620, 2016.