Asymptotic post-selection inference after model selection by the Akaike information criterion.

Gerda Claeskens

ORSTAT and Leuven Statistics Research Center, KU Leuven, Belgium

WHOA-PSI-2017, St. Louis, 12–14 August 2017
Confidence intervals post-AIC selection

Joint work with Ali Charkhi (KU Leuven)
Model selection by AIC

- Model selection by Akaike’s information criterion

\[ \text{AIC}(M_k) = -2 \log \text{Likelihood}_{M_k}(\hat{\beta}_{M_k}) + 2p_{M_k} \]

Select model \( \hat{M}_{\text{aic}} \) for which AIC(\( \hat{M}_{\text{aic}} \)) is the smallest amongst AIC(\( M_k \)), \( k = 1, \ldots, K \)

- Estimation in the selected model:

\[ \mu(\hat{\beta}_{\hat{M}_{\text{aic}}}) = \sum_{k=1}^{K} \mu(\hat{\beta}_{M_k}) \cdot I(M_k = \hat{M}_{\text{aic}}) \]

Selection is a special case of model averaging:

- Weights are only 0 or 1
- Weights are random, since data-driven.
- Randomness of weights influences further inference
Confidence regions after AIC selection

In selected model \( \hat{M}_{aic} \), estimate focus \( \mu \) by

\[
\hat{\mu}(\hat{M}_{aic}) = \sum_{k=0}^{K} \hat{\mu}(M_k) \cdot I(M_k = \hat{M}_{aic}).
\]

Confidence regions for \( \mu \) after selection?
Naive interval: pretend that the model has been given beforehand. Ignore randomness \( \hat{M}_{aic} = M_{aic} \).

\[
\hat{\mu}(M_{aic}) \pm 1.96 \cdot \text{standard error}(\hat{\mu}(M_{aic}))
\]

Frequentist model averaging necessary to overcome low coverage of using naive method after AIC-selection.

(Hjort and Claeskens, 2003)
Some recent literature on correct inference


Efron (2014): calculate the variance of the bagging estimator ⇒ not based on the selected model for the original data.


Schneider (2016): coverage of intervals based on thresholding estimators in high-dimensional linear regression models.
Post-AIC-selection in nested models

Classical (easy) setting to start off with:

- $\mathcal{M}_{\text{nest}} = \{M_k; M_1 \subset M_2 \subset \ldots \subset M_K\}$
- Maximum likelihood estimation in each model $\hat{\theta}_k = (\hat{\theta}_k^t, 0_{K-k}^t)$.
- Minimal true model $M_{p_0} \in \mathcal{M}_{\text{nest}}$:
  - Models with indices $k < p_0$ are under-parametrized,
  - Models with $k > p_0$ are over-parametrized.
- $\text{AIC}(M_k) = -2\ell_n(\hat{\theta}_k) + 2(a + k)$

Select $\hat{p}_{\text{aic}}$ if and only if $\text{AIC}(M_{\hat{p}_{\text{aic}}}) \leq \text{AIC}(M_k)$ for all $k = 1, \ldots, K$
Asymptotics of AIC selection

AIC overselects (Nested models: Woodroofe 1982, arc-sine laws) for $n \to \infty$, $\hat{p}_{aic} \in \{p_0, p_0 + 1, \ldots, K\}$,

Minimize AIC $\iff$ maximize $\text{AIC}^*$

$$\text{AIC}^*(M_k) = 2\{\ell_n(\hat{\theta}_{(k)}) - \ell_n(\vartheta)\} - 2k = 2\ell^*_n,k - 2k.$$ 

Joint convergence

$$2(\ell^*_{n,p_0}, \ldots, \ell^*_{n,K}) \xrightarrow{d} (\sum_{i=1}^{a+p_0} Z_i^2, \ldots, \sum_{i=1}^{a+K} Z_i^2)$$

with $Z_1, \ldots, Z_{a+K} \overset{i.i.d.}{\sim} N(0, 1)$
AIC selection region – Example

Three models: $M_1 : \{\theta_1\}; M_2 : \{\theta_1, \theta_2\}; M_3 : \{\theta_1, \theta_2, \theta_3\}$

$a = 1, K = 2$, take $M_1$ smallest true model, $M_3$ selected, $\hat{p}_{aic} = 2$

\[
\left\{ \begin{array}{l}
\text{AIC}^*(M_3) > \text{AIC}^*(M_2) \\
\text{AIC}^*(M_3) > \text{AIC}^*(M_1)
\end{array} \right.
\]

Asymptotic representation:

\[
\left\{ \begin{array}{l}
Z_1^2 + Z_2^2 + Z_3^2 - 6 > Z_1^2 + Z_2^2 - 4 \\
Z_1^2 + Z_2^2 + Z_3^2 - 6 > Z_1^2 - 2
\end{array} \right.
\]

$A_2(M_{nest}) = \{z \in \mathbb{R}^3 : z_3^2 > 2, z_2^2 + z_3^2 > 4\}.$
Asymptotic distribution

\( J_p \): Fisher information for model \( M_p \).
\( \tilde{\nu}(p) = (\nu_1, \ldots, \nu_{a+p})^t \): sub-vector of \( \nu = (\nu_1, \ldots, \nu_{a+K})^t \) for model \( M_p \).
\( \vartheta \) true value

Post-AIC distribution for nested models

For \( M_{\text{nest}} \) with \( p_0 \) denoting the true model order

\[
F_p(t) = \lim_{n \to \infty} P(n^{1/2}(\hat{\theta}(p) - \vartheta) \leq t \mid \hat{p}_{\text{aic}} = p, M_{\text{nest}}) \\
= P\{J_p^{-1/2}\tilde{Z}(p) \leq \tilde{t}(p) \mid \tilde{Z}(p) \in A_p^{(s)}(M_{\text{nest}})\} \cdot I(t \in T_p),
\]

where \( A_p^{(s)}(M_{\text{nest}}) = \{\tilde{z}(p) \in \mathbb{R}^{a+p} : \bigcap_{j=p_0+1,\ldots,p} [\sum_{i=j}^p (z_{a+i}^2 - 2) > 0] \} \)
and \( T_p = \mathbb{R}^{a+p} \times (\mathbb{R}^+)^{K-p} \).

Simplified region for nested models due to independence of \( \tilde{Z}(p) \) and \( (Z_{p+1}, \ldots, Z_K) \).
Asymptotic density

**Truncated multivariate normal density**

Denote $\phi_p(\cdot|A; \Sigma)$ the density of $\Sigma^{-1/2}\widetilde{Z}(p)$ where $\widetilde{Z}(p) \sim N_{a+p}(0, I_{a+p})$ is truncated such that $\widetilde{Z}(p) \in A$.

**Corollary: Post-AIC density for nested models**

The limiting density function of $n^{1/2}(\hat{\theta}(\hat{p}_{aic}) - \vartheta)$ conditional on the AIC-selection with $\hat{p}_{aic} = p$ from the set of nested models $\mathcal{M}_{nest}$ is

$$f_p(t) = \phi_p(\tilde{t}(p)|A_p(s)(\mathcal{M}_{nest}); J_p^{-1}) \cdot I(t \in T_p).$$

When the true model is selected, $\hat{p}_{aic} = p_0$,

$$f_{p_0}(t) = \phi_{p_0}(\tilde{t}(p_0)) \cdot I(t \in T_p)$$
Exact calculations for nested models

True value for parameters: \( \vartheta = (\vartheta_1, 0, 0)^t \leftrightarrow M_1 \) true.
Models set: \( \mathcal{M}_{nest} = \{M_1, M_2, M_3\} \).
Selected model: \( M_3 \Rightarrow A_3 = \{(z_1, z_2, z_3): z_3^2 > 2, z_3^2 + z_2^3 > 4\} \).
Selection probability: \( P(Z \in A_3) = 0.08 \).

\[ J^{-1/2}(\vartheta): \]

\[
\begin{pmatrix}
1.0 & 0 & 0 \\
0 & 2.0 & 0 \\
0 & 0 & 2.0
\end{pmatrix}, \quad
\begin{pmatrix}
1.0 & 0 & 0 \\
0 & 2.0 & 0.5 \\
0 & 0.5 & 2.0
\end{pmatrix}, \quad
\begin{pmatrix}
1.0 & 0.9 & 0.9 \\
0.9 & 2.0 & 0.5 \\
0.9 & 0.5 & 2.0
\end{pmatrix}.
\]
Exact calculations for nested models

Exact marginal asymptotic distributions of $\sqrt{n}(\hat{\theta}_3 - \vartheta_3)$ when $M_1$ is true model and $M_3$ is selected, for different Fisher information matrices.
Simulation based inference

Sample from a multivariate normal distribution subject to quadratic constraints. (Pakman and Paninski 2014)

(a) (b) (c)

<table>
<thead>
<tr>
<th>Simulated quantile coverage</th>
<th>$\chi^2$ quantile coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>12.17</td>
<td>0.949</td>
</tr>
<tr>
<td>7.81</td>
<td>0.826</td>
</tr>
</tbody>
</table>
Inference depends on the set of models

Nested models:
\( M_1 : \{ \theta_1 \} ; \ M_2 : \{ \theta_1, \theta_2 \} ; \ M_3 : \{ \theta_1, \theta_2, \theta_3 \} \)

AIC selects \( M_3 \):
\( \text{AIC}(M_3) > \text{AIC}(M_1) \) and \( \text{AIC}(M_3) > \text{AIC}(M_2) \)

\( A_M(M_{nest}) = \{ z \in \mathbb{R}^3 : z_3^2 > 2, z_2^2 + z_3^2 > 4 \} \).

\( \mathcal{M}_{all} = \text{All subsets of } M_3 : \{ \theta_1, \theta_2, \theta_3 \} \)

AIC selects \( M_3 \):
\( \text{AIC}(M_3) > \text{AIC}(M') \) for all models \( M' \neq M_3 \in \mathcal{M} \)

\( A_M(M_{all}) = \{ z \in \mathbb{R}^3 : z_2^2 > 2, z_3^2 > 2, z_2^2 + z_3^2 > 4 \} \)

Different area, different conditioning

\[ \Rightarrow \text{different distribution} \]
\[ \Rightarrow \text{different quantiles} \]
\[ \Rightarrow \text{different confidence intervals} \]
General likelihood models, any model set

For $M_{aic} = M \in \mathcal{M}_O$, overparametrized model, the confidence region

$$C(q_\alpha) = \left\{ \theta \in \mathbb{R}^{a+K} : n(\hat{\theta}_M - \tilde{\theta}_M)^t J_M (\hat{\theta}_M - \tilde{\theta}_M) \leq q_\alpha \right\},$$

where $q_\alpha$ is determined by solving

$$P \left( \left\{ \sum_{i \in M} Z_i^2 \leq q_\alpha \right\} \cap Z \in A_M (\mathcal{M}_{arb} \cap \mathcal{M}_O) \right) \frac{P(Z \in A_M (\mathcal{M}_{arb} \cap \mathcal{M}_O))}{P(Z \in A_M (\mathcal{M}_{arb} \cap \mathcal{M}_O))} = 1 - \alpha.$$

Uniform result

Under some assumptions, incl. compact parameter space.

$$\lim_{n \to \infty} \inf_{\theta \in \Theta} P_{\theta}\{ \theta \in C(q_\alpha) \mid M_{aic} \in \mathcal{M}_O \} = 1 - \alpha.$$

When $A_M (\mathcal{M}_{arb})$ replaces $A_M (\mathcal{M}_{arb} \cap \mathcal{M}_O)$ to obtain $\tilde{q}_\alpha$, 

$$\lim_{n \to \infty} \inf_{\theta \in \Theta} P\{ \theta \in C(\tilde{q}_\alpha) \mid M_{aic} \in \mathcal{M}_O \} \geq 1 - \alpha.$$
About the uniformity result

- Limitation of AIC: selection of an overspecified model does not happen in a uniform way (Leeb, Pötscher, 2003).

- An overspecified model is needed. Hence result cannot be strengthened for AIC under this setup.

- For underparametrized model: need to work with pseudo-true values.

- ‘Targets’ differ from model to model.
All likelihood models misspecified

- Triangular array of observations \( \{Y_{ni}; i = 1, \ldots, n, \text{ and } n \in \mathbb{N}_0\} \);
- True distribution of \((Y_{n1}, \ldots, Y_{nn})\) is \(G_n\);
- Estimators maximizing \( \prod_{i=1}^{n} f_{ji}(y_i; \theta_j) \) converge to pseudo-true values \( \vartheta^*_n(M_j), j = 1, \ldots, K \);
- When there is an estimator \( \hat{\Sigma} \) of \( \Sigma \) such that

  \[
  \lim_{n \to \infty} \sup_{G_n \in \mathcal{G}_n} P_{G_n}(\|\hat{\Sigma}_n - \Sigma\| > \varepsilon) = 0,
  \]

then, with \( Z_{m'} \sim N_{m'}(0, I_{m'}) \)

  \[
  \lim_{n \to \infty} \sup_{G_n \in \mathcal{G}_n} \sup_{t \in \mathbb{R}^{m'}} |P(\hat{\Sigma}_n^{-1/2} n^{-1/2}(\hat{\theta}_{n,M} - \vartheta^*_n,M) \leq t) - P(Z_{m'} \leq t)| = 0.
  \]

- White (1994) contains conditions. Sandwich estimator might overestimate \( \rightarrow \) conservative confidence intervals.
Assume (i) that there is a smallest model nested in all other models, (ii) similarity assumption of Vuong (1989), then construct selection region $A_{\text{M} \text{aic}}(\mathcal{M})$ similar as before

### Uniform result under likelihood misspecification

$$
\lim_{n \to \infty} \sup_{G_n \in \mathcal{G}_n} \sup_{t \in \mathbb{R}^{|M_{\text{aic}}|}} \left| P\left( n^{1/2} \{ \hat{\theta}(M_{\text{aic}}) - \vartheta^*(M_{\text{aic}}) \} \leq t \mid M_{\text{aic}} \right) \right.
$$

$$
- P \left( \Sigma_{M_{\text{aic}}}^{1/2} Z \leq t \mid A_{M_{\text{aic}}}(\mathcal{M}) \right) \left| = 0 \right.
$$

Define the set

$$
C^*(q_\alpha) = \{ \theta \in \mathbb{R}^{|M_{\text{aic}}|} : n \{ \hat{\theta}(M_{\text{aic}}) - \theta(M_{\text{aic}}) \}^t \Sigma_{M_{\text{aic}}} \vartheta_{M_{\text{aic}}}^{-1} \{ \hat{\theta}(M_{\text{aic}}) - \theta(M_{\text{aic}}) \} \leq q_\alpha \},
$$

Then

$$
\lim_{n \to \infty} \sup_{G_n \in \mathcal{G}_n} \sup_{\alpha \in [0,1]} \left| P_{G_n} \{ \vartheta^*(M_{\text{aic}}) \in C^*(q_\alpha) \mid M_{\text{aic}} \} - (1 - \alpha) \right| = 0.
$$
Simulation study – Linear Regression

Model to generate data

\[ Y_i = \sum_{j=1}^{10} \theta_j x_{ji} + \varepsilon_i, \quad i = 1, \ldots, n, \]

- \( \varepsilon_i \sim N(0, 1) \)
- \( \boldsymbol{\theta}^t = (\theta_1, \ldots, \theta_{10})^t = (2.25, -1.1, 2.43, -2.24, 2.5, 0.5^t). \)
- \( x_{1i} = 1 \) and \( (x_{2i}, \ldots, x_{10, i})^t \sim N(0_9, \Omega) \) where \( \Omega_{ii} = 1 \) and \( \Omega_{ij} = 0.25, \ j \neq i. \)

Selection by AIC from \( \zeta_i^i_{all} \): first \( i \) parameters are present in all models, all combinations of other parameters.

Focus model: \( M = (\theta_1, \ldots, \theta_6, \theta_8) \): Run simulation until \( M \) is selected 3000 times.

Report average confidence intervals over 3000 runs and coverage probabilities.
CI for regression parameter

(Berk, Brown, Buja, Zhang, Zhao 2013 AoS)

<table>
<thead>
<tr>
<th>n</th>
<th>method</th>
<th>$\theta_j$</th>
<th>$\zeta^3_{all}$</th>
<th>Cov.</th>
<th>$\zeta^6_{all}$</th>
<th>Cov.</th>
</tr>
</thead>
<tbody>
<tr>
<td>PostAIC</td>
<td>$\theta_4$</td>
<td>[-2.54, -1.94]</td>
<td>0.99</td>
<td>[-2.46, -2.02]</td>
<td>0.94</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\theta_6$</td>
<td>[-0.30, 0.31]</td>
<td>0.95</td>
<td>[-0.22, 0.22]</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\theta_8$</td>
<td>[-0.30, 0.31]</td>
<td>0.95</td>
<td>[-0.29, 0.30]</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>PoSI</td>
<td>$\theta_4$</td>
<td>[-2.58, -1.90]</td>
<td>1.00</td>
<td>[-2.54, -1.94]</td>
<td>0.99</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>$\theta_6$</td>
<td>[-0.33, 0.34]</td>
<td>0.98</td>
<td>[-0.30, 0.30]</td>
<td>0.99</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\theta_8$</td>
<td>[-0.34, 0.34]</td>
<td>0.98</td>
<td>[-0.29, 0.31]</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>Naive</td>
<td>$\theta_4$</td>
<td>[-2.46, -2.02]</td>
<td>0.93</td>
<td>[-2.46, -2.02]</td>
<td>0.93</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\theta_6$</td>
<td>[-0.22, 0.22]</td>
<td>0.66</td>
<td>[-0.22, 0.22]</td>
<td>0.94</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\theta_8$</td>
<td>[-0.22, 0.22]</td>
<td>0.66</td>
<td>[-0.22, 0.23]</td>
<td>0.69</td>
<td></td>
</tr>
</tbody>
</table>

$\zeta^3_{all}$: More conservative for $\theta_4$.

$\zeta^6_{all}$: Almost exact for all parameters.
CI for a linear combination of the parameters $x^t \theta$

PoSlp (Bachoc, Leeb, Pötscher, 2015)
Smoothed bootstrap CI (Efron, 2014)

<table>
<thead>
<tr>
<th>$\sigma$</th>
<th>$n$</th>
<th>method</th>
<th>$\zeta^3_{all}$</th>
<th>cov.</th>
<th>$\zeta^5_{all}$</th>
<th>cov.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>30</td>
<td>PostAIC</td>
<td>3.11</td>
<td>0.97</td>
<td>2.61</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bootstrap</td>
<td>3.67</td>
<td>0.92</td>
<td>3.32</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PoSlp</td>
<td>4.38</td>
<td>1.00</td>
<td>4.39</td>
<td>1.00</td>
</tr>
<tr>
<td>100</td>
<td></td>
<td>PostAIC</td>
<td>1.42</td>
<td>0.98</td>
<td>1.17</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bootstrap</td>
<td>1.25</td>
<td>0.94</td>
<td>1.25</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PoSlp</td>
<td>1.83</td>
<td>1.00</td>
<td>1.83</td>
<td>1.00</td>
</tr>
<tr>
<td>3</td>
<td>30</td>
<td>PostAIC</td>
<td>11.76</td>
<td>0.98</td>
<td>7.82</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bootstrap</td>
<td>11.46</td>
<td>0.92</td>
<td>9.95</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PoSlp</td>
<td>12.65</td>
<td>0.99</td>
<td>13.16</td>
<td>1.00</td>
</tr>
<tr>
<td>100</td>
<td></td>
<td>PostAIC</td>
<td>4.25</td>
<td>0.98</td>
<td>3.50</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bootstrap</td>
<td>3.77</td>
<td>0.94</td>
<td>3.74</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PoSlp</td>
<td>5.47</td>
<td>1.00</td>
<td>5.48</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Extensions to other criteria

AIC-like criteria, overselection is exploited.

- **Takeuchi’s information criterion** (1976)

  \[
  \text{TIC}(M) = -2\ell_n(\hat{\theta}(M)) + 2\text{tr}\{Q_M(\vartheta^*)^{-1}J_M(\vartheta^*)\}.
  \]

  Replace \(|M|\) with \(\text{tr}\{Q_M(\vartheta^*)^{-1}J_M(\vartheta^*)\}\)

- **Generalized information criterion** (Konishi, Kitagawa 1996)

  \[
  \text{GIC}(M) = -2\ell_n(\hat{\theta}(M)) + \frac{2}{n} \sum_{i=1}^{n} \text{tr}\{\text{Infl}(Y_i)\frac{\partial}{\partial \theta_M^t} \log f(Y_i; \hat{\theta}_M)\}.
  \]

- **Mallows’s \(C_p\)** for linear regression (1973)

  \[
  C_p(M) = \hat{\sigma}^{-2} \hat{\sigma}^2(M) + 2|M| - n
  \]

  When \(n \to \infty\), \(C_p(M) - C_p(M^*) \sim \chi^2_q/q + 2q\) where \(q = |M^*| - |M|\)
Take home messages

▶ Asymptotic AIC selection regions are formed by quadratic functions.
▶ Need overparametrization when a true model exists.
▶ Stronger uniform results under misspecification (pseudo-true parameter values).
▶ Think carefully about the sets of models to include in a search.

Thank you!