

Valid confidence intervals for post-model-selection predictors

François Bachoc, Hannes Leeb et Benedikt M. Pötscher

University Paul Sabatier, Toulouse
University of Vienna

Workshop on post-model selection
Leuven, August 22-23, 2016

Talk outline

- 1 Introduction and overview
- 2 Confidence intervals for the design-dependent target
- 3 Confidence intervals for the design-independent target
- 4 Simulation study

Data generating process

Location model

$$\mathbf{Y} = \boldsymbol{\mu} + \mathbf{U}$$

- \mathbf{Y} of size $n \times 1$: observation vector
- $\boldsymbol{\mu}$ of size $n \times 1$: unknown mean vector
- $\mathbf{U} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$
- σ^2 unknown

\implies Working distribution $P_{n, \boldsymbol{\mu}, \sigma}$

Linear submodels

Consider a design matrix \mathbf{X} of size $n \times p$

- $p < n$ or $p \geq n$

Linear submodels

Subsets $M \subset \{1, \dots, p\}$ of the columns of \mathbf{X} . Approximating μ by

$$\mathbf{X}[M]\mathbf{v}$$

- M of cardinality $m \leq n$
- $\mathbf{X}[M]$ of size $n \times m$: only the columns of \mathbf{X} that are in M
- $\mathbf{X}[M]$ full rank
- \mathbf{v} of size $m \times 1$: needs to be selected/estimated

Restricted least square estimator

$$\hat{\beta}_M = (\mathbf{X}'[M]\mathbf{X}[M])^{-1} \mathbf{X}'[M]\mathbf{Y}$$

Regression coefficients of interest

Two main classes of regression coefficients

- **The 'standard' target** Assume here that

$$\boldsymbol{\mu} = \mathbf{X}[M^*]\boldsymbol{\beta}_0$$

with M^* of size $m^* \leq n$ and $\boldsymbol{\beta}_0$ of size m^*

Then $\boldsymbol{\beta}_0$ and $\boldsymbol{\mu}$ are targets of inference

- **The 'projection-based' target** No assumption of $\boldsymbol{\mu}$

Let for M of size $m \leq n$

$$\boldsymbol{\beta}_M^{(n)} = \underset{\mathbf{v}}{\operatorname{argmin}} \|\boldsymbol{\mu} - \mathbf{X}[M]\mathbf{v}\|$$

$$\boldsymbol{\beta}_M^{(n)} = (\mathbf{X}'[M]\mathbf{X}[M])^{-1} \mathbf{X}'[M]\boldsymbol{\mu}$$

Then $\boldsymbol{\beta}_M^{(n)}$ is a target of inference

Model selection

Model selection procedure

Data-driven selection of the model with $\hat{M}(\mathbf{Y}) = \hat{M}$
 $\hat{M} \in \mathcal{M}$ with \mathcal{M} the universe of possible models (fixed and known)

Ex. : sequential testing, AIC, BIC, LASSO

The presence of model selection makes it **more difficult** to make inference on $\beta_{\hat{M}}^{(n)}$ or β_0 based on $\hat{\beta}_{\hat{M}}$

See e.g. [Leeb and Poëtscher 05, 06, 12](#)

This is what we call the **post-model-selection inference** problem

Post-model-selection in the literature

- In [Van der Geer et al. 2014](#), confidence intervals for β_0 are constructed based on the LASSO estimator
- In [Lee et al. 2016](#), confidence intervals for $\beta_{\hat{M}}^{(n)}$ are constructed when \hat{M} is the LASSO model selector
- In [Berk et al 2013, annals of statistics](#), the target for inference is $\beta_{\hat{M}}^{(n)}$ and \hat{M} can be [any](#) model selection procedure
 - Model selector \hat{M} is "imposed"
 - Objective : best coefficients in this imposed model

Overview of our contributions

In this paper

- ✎ **F. Bachoc, H. Leeb, B.M. Pötscher. Valid confidence intervals for post-model-selection predictors,**
<http://arxiv.org/abs/1412.4605>

we extend the work of [Berk et al. 2013](#) to predictors

- that is, the target is $\mathbf{x}_0[\hat{M}]'\beta_M^{(n)}$ instead of $\beta_M^{(n)}$

($\mathbf{x}_0[\hat{M}]$ is obtained from \mathbf{x}_0 by keeping components in \hat{M})

Two main contributions

- When $p > n$ is allowed, extension of the procedure of Berk et al. to prediction and large p analysis of the confidence intervals
(design-dependent target)
- When $p \leq n$, definition of a more beneficial target $\mathbf{x}_0[\hat{M}]'\beta_M^{(*)}$
(design-independent target) and large n analysis

- 1 Introduction and overview
- 2 Confidence intervals for the design-dependent target**
- 3 Confidence intervals for the design-independent target
- 4 Simulation study

Predictors

Let

$$y_0 = \mu_0 + u_0$$

- $u_0 \sim \mathcal{N}(0, \sigma^2)$

Let \mathbf{x}_0 be a $p \times 1$ vector

We consider the **design-dependent non-standard target**

$$\mathbf{x}'_0[\hat{M}]\beta_{\hat{M}}^{(n)}$$

Then, we have when \mathbf{x}_0 follows the empirical distribution given by the lines of \mathbf{X} ,

$$\mathbb{E} \left(\left[y_0 - \mathbf{x}'_0[\hat{M}]\beta_{\hat{M}}^{(n)} \right]^2 \right) \leq \mathbb{E} \left(\left[y_0 - \mathbf{x}'_0[\hat{M}]\mathbf{v}(\mathbf{Y}) \right]^2 \right),$$

for any function $\mathbf{v}(\mathbf{Y}) \in \mathbb{R}^{|\hat{M}|}$

Estimation of σ

Let \mathbf{P}_X denote the orthogonal projection on the column space of \mathbf{X}

Assumption

We have available an (observable) random variable $\hat{\sigma}^2$ that is independent of $\mathbf{P}_X \mathbf{Y}$ and that is distributed as σ^2/r times a chi-square distributed random variable with r degrees of freedom ($1 \leq r \leq \infty$)

- Same assumption as in Berk et al. 2013
- Satisfied if $p \leq n$ and $\boldsymbol{\mu} \in \text{span}(\mathbf{X})$
- Otherwise not innocuous !
- Worthy of further research

Confidence intervals

Let a nominal level $1 - \alpha \in (0, 1)$ be fixed

We consider confidence intervals for $\mathbf{x}'_0[\hat{M}]\beta_M^{(n)}$ of the form

$$CI = \mathbf{x}'_0[\hat{M}]\hat{\beta}_{\hat{M}} \pm K \|\mathbf{s}_{\hat{M}}\| \hat{\sigma},$$

with

$$\mathbf{s}'_M = \mathbf{x}'_0[M] (\mathbf{X}'[M]\mathbf{X}[M])^{-1} \mathbf{X}'[M]$$

Interpretation

- "Constant" K does not depend on \mathbf{Y} (but on \mathbf{X} , \mathbf{x}_0 , \hat{M})
- For **fixed** M ,

$$\mathbf{x}'_0[M]\hat{\beta}_M - \mathbf{x}'_0[M]\beta_M^{(n)} \sim \mathcal{N}(0, \|\mathbf{s}_M\|^2 \sigma^2)$$

- Thus, $K_{naive} = q_{S,r,1-\alpha/2}$ (Student quantile) is valid when M is deterministic
- When \hat{M} is random, K needs to be larger (e.g. [Leeb et al. 2015, Statistical Science](#))

⇒ **Main issue** : choosing K

The construction of Berk et al.

Observe that

$$\mathbf{x}'_0[\hat{M}]\hat{\beta}_{\hat{M}} - \mathbf{x}'_0[\hat{M}]\beta_{\hat{M}}^{(n)} = \mathbf{s}'_{\hat{M}}(\mathbf{Y} - \mu)$$

Then, we have

$$\left| \frac{\mathbf{s}'_{\hat{M}}}{\|\mathbf{s}_{\hat{M}}\|\hat{\sigma}}(\mathbf{Y} - \mu) \right| \leq \max_{M \subseteq \{1, \dots, p\}} \left| \frac{\mathbf{s}'_M}{\|\mathbf{s}_M\|\hat{\sigma}}(\mathbf{Y} - \mu) \right|$$

Distribution of the upper-bound **independent** of $\mu, \sigma \implies$ let K_1 be its $(1 - \alpha)$ quantile

The CI given by K_1 satisfies

$$\inf_{\mu \in \mathbb{R}^n, \sigma > 0} P_{n, \mu, \sigma} \left(\mathbf{x}'_0[\hat{M}]\beta_{\hat{M}}^{(n)} \in CI \right) \geq 1 - \alpha$$

\implies **Uniformly valid** confidence interval

Computing K_1

Let $\bar{\mathbf{s}}_M = \mathbf{s}_M / \|\mathbf{s}_M\|$

Since $\bar{\mathbf{s}}_M$ belongs to the column space of \mathbf{X} for every $M \in \mathcal{M}$, we have

$$\begin{aligned} & P_{n, \mu, \sigma} \left(\max_{M \in \mathcal{M}} |\bar{\mathbf{s}}_M' (\mathbf{Y} - \boldsymbol{\mu})| / \hat{\sigma} > t \right) \\ &= P_{n, \mu, \sigma} \left(\max_{M \in \mathcal{M}} |\bar{\mathbf{s}}_M' \mathbf{P}_X (\mathbf{Y} - \boldsymbol{\mu}) / \|\mathbf{P}_X (\mathbf{Y} - \boldsymbol{\mu})\| \| \mathbf{P}_X (\mathbf{Y} - \boldsymbol{\mu}) \| > (\hat{\sigma} / \|\mathbf{P}_X (\mathbf{Y} - \boldsymbol{\mu})\|) t \right) \\ &= \Pr \left(\max_{M \in \mathcal{M}} |\bar{\mathbf{s}}_M' \mathbf{V}| > t/G \right) \end{aligned}$$

where

- $d = \text{rank}(\mathbf{X})$
- \mathbf{V} is uniformly distributed on the unit sphere of the column space of \mathbf{X}
- G^2/d follows an F-distribution with $(d-r)$ -degrees of freedom

Computing K_1 : algorithm

We need to compute the K so that

$$\alpha = \Pr \left(\max_{M \in \mathcal{M}} |\bar{\mathbf{s}}'_M \mathbf{V}| > K/G \right) = \mathbb{E}_V \left(1 - F_{d,r} \left(K^2 / \left\{ \max_{M \in \mathcal{M}} |\bar{\mathbf{s}}'_M \mathbf{V}|^2 d \right\} \right) \right)$$

with $F_{d,r}$ the c.d.f. of the Fisher (d, r) distribution

Algorithm

Choose $l \in \mathbb{N}$ and generate independent identically distributed random vectors $\mathbf{V}_1, \dots, \mathbf{V}_l$, where each \mathbf{V}_i is uniformly distributed on the unit sphere of the column space of \mathbf{X} . Calculate the quantities $c_i = \max_{M \in \mathcal{M}} |\bar{\mathbf{s}}'_M \mathbf{V}_i|$. A numerical approximation to K_1 is then obtained by searching for that value of K that solves

$$\frac{1}{l} \sum_{i=1}^l F_{d,r} \left(\frac{K^2}{c_i^2 d} \right) = 1 - \alpha.$$

- Complexity in 2^p if \mathcal{M} consists in all subsets of $\{1, \dots, p\}$
- In practice, one hour for $p = 20$
- For larger p , one should use upper-bounds of K_1 , cf below

Issues when \mathbf{x}_0 is partially observed

The constant K_1 depends on all the components of \mathbf{x}_0

It can happen that only $\mathbf{x}_0[\hat{M}]$ is observed

- model selection for cost reason

We hence construct other constants so that

$$K_1 \leq K_2 \leq K_3 \leq K_4$$

(The CIs given by K_2, K_3, K_4 are hence universally valid)

K_2, K_3, K_4 depend only on $\mathbf{x}_0[\hat{M}]$

Remark : K_4 is introduced by Berk et al.

K_2

Let

$$K_2(\mathbf{x}_0[\hat{M}], \hat{M}) = \sup \left\{ K_1(\mathbf{x}) : \mathbf{x}[\hat{M}] = \mathbf{x}_0[\hat{M}] \right\},$$

- Maximizing K_1 over the unobserved components of \mathbf{x}_0 (those which are not in the submodel \hat{M})
- Optimally small
- (too) costly to compute (stochastic optimization of K_1)

K_3 and K_4

Recall that $K_1(x_0)$ is the K so that

$$\Pr \left(\max_{M \in \mathcal{M}} |\bar{\mathbf{s}}'_M \mathbf{V}| > K/G \right) = \alpha$$

and that the problem is that $\bar{\mathbf{s}}_M$ is not observed when M is not included in \hat{M}
 Union bound :

$$\begin{aligned} \Pr \left(\max_{M \in \mathcal{M}} |\bar{\mathbf{s}}'_M \mathbf{V}| > K/G \right) &= \mathbb{E}_G \left\{ \Pr \left(\max_{M \in \mathcal{M}} |\bar{\mathbf{s}}'_M \mathbf{V}| > K/G \right) \right\} \\ &\leq \mathbb{E}_G \left\{ \Pr \left(\max_{M \in \mathcal{M}, M \subseteq \hat{M}} |\bar{\mathbf{s}}'_M \mathbf{V}| > \frac{K}{G} \right) + c(\hat{M}, \mathcal{M}) \left\{ 1 - F_{beta, 1/2, (d-1)/2}(K^2/G^2) \right\} \right\} \end{aligned} \quad (1)$$

$$\leq \mathbb{E}_G \left\{ |\mathcal{M}| \left\{ 1 - F_{beta, 1/2, (d-1)/2}(K^2/G^2) \right\} \right\} \quad (2)$$

- $\Pr \left(|\bar{\mathbf{s}}'_M \mathbf{V}| > \frac{K}{G} \right) = 1 - F_{beta, 1/2, (d-1)/2}(K^2/G^2)$
- $c(\hat{M}, \mathcal{M})$ is the number of models $M \in \mathcal{M}$ so that $M \not\subseteq \hat{M}$
- $K_3(x_0[\hat{M}], \hat{M})$ and K_4 are the values of K so that (1) and (2) equal α

K_3 and K_4

- K_3 has the same computational cost as K_1
- K_4 is cheap to compute
- K_4 was proposed in [Berk et al. 2013](#)

Large p analysis of K_1

- K_1 depends on \mathbf{x}_0 and \mathbf{X} , and it does not seem easy to provide a systematic large p analysis, for any \mathbf{X}, \mathbf{x}_0
- When $\mathbf{x}_0 = \mathbf{e}_i$ (base vector), Berk et al. 2013 show that
 - When \mathbf{X} has orthogonal columns, K_1 has rate $\sqrt{\log(p)}$
 - There exists sequences of \mathbf{X} so that K_1 has rate \sqrt{p}

We show

Proposition

Let \mathcal{M} be the power set of $\{1, \dots, p\}$. Let $\alpha, 0 < \alpha < 1$, be given

(a) Let \mathbf{X} have orthogonal columns. There exist a sequence of vectors \mathbf{x}_0 such that K_1 satisfies

$$\liminf_{p \rightarrow \infty} K_1(\mathbf{x}_0) / \sqrt{p} \geq 0.6363$$

(b) Let $\gamma \in [0, 1)$ be given. Then $K_2(\mathbf{x}_0[M], M)$ satisfies

$$\liminf_{p \rightarrow \infty} \inf_{\mathbf{x}_0 \in \mathbb{R}^p} \inf_{\mathbf{X} \in \mathcal{X}(p)} \inf_{M \in \mathcal{M}, |M| \leq \gamma p} K_2(\mathbf{x}_0[M], M) / \sqrt{p} \geq 0.6363 \sqrt{1 - \gamma},$$

where $\mathcal{X}(p) = \bigcup_{n \geq p} \{\mathbf{X} : \mathbf{X} \text{ is } n \times p \text{ with non-zero orthogonal columns}\}$

Large p analysis of K_3 and K_4

Using results from [Berk et al. 2013](#), [Zhang 2015](#), we show

Proposition

Assume that $\mathcal{M} = \mathcal{M}_p$ satisfies

- (i) $\bigcup \{M : M \in \mathcal{M}\} = \{1, \dots, p\}$
- (ii) $c(M, \mathcal{M}) \geq \tau |\mathcal{M}|$ for every $M \in \mathcal{M}$ with $M \neq \{1, \dots, p\}$

Let $X_{n,p}(\mathcal{M})$ denote the set of all $n \times p$ matrices of rank $\min(n, p)$ with the property that $X[M]$ has full column-rank for every $\emptyset \neq M \in \mathcal{M}$

Then we have

$$\lim_{p \rightarrow \infty} \sup_{M \in \mathcal{M}, M \neq \{1, \dots, p\}} \sup_{x_0 \in \mathbb{R}^p} \sup_{X \in X_{n(p), p}(\mathcal{M})} |1 - (K_3(x_0[M], M)/K_4)| = 0$$

Furthermore

$$K_4 / \sqrt{\min(n(p), p) \left(1 - |\mathcal{M}|^{-2/(\min(n(p), p) - 1)}\right)} \rightarrow 1$$

as $p \rightarrow \infty$

Large p analysis of K_3 and K_4

K_4 gets smaller when \mathcal{M} gets smaller :

- K_4 always asymptotically no larger than $\sqrt{\min(n, p)}$
- When $p > n$ and $\mathcal{M} = \{M \subseteq \{1, \dots, p\}; |M| \leq an^b\}$ with fixed a and $0 \leq b < 1$, we have

$$K_4 = O(n^{b/2} \sqrt{\log(n)})$$

- 1 Introduction and overview
- 2 Confidence intervals for the design-dependent target
- 3 Confidence intervals for the design-independent target**
- 4 Simulation study

Settings for this section

Well-specified linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}$$

- \mathbf{Y} of size $n \times 1$
- \mathbf{X} of size $n \times p$
- $\boldsymbol{\beta}$ of size $p \times 1$: fixed and unknown
- $\mathbf{U} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$
- $0 < \sigma < \infty$ fixed and unknown
- p fixed, $n \rightarrow \infty$

\implies Working distribution $P_{n,\beta,\sigma}$

Least square estimator :

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y}$$

Standard variance estimator :

$$\hat{\sigma}^2 = \frac{1}{n-p} \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2$$

Design-independent non-standard target

Issue : The target $\mathbf{x}'_0[\hat{M}]\beta_{\hat{M}}^{(n)}$ depends on \mathbf{X} but is a predictor of y_0 from \mathbf{x}_0

Issue is solved when lines of \mathbf{X} and \mathbf{x}'_0 are realizations from the same distribution \mathcal{L}

Let, for $\mathbf{x}' \sim \mathcal{L}$, $\Sigma = \mathbb{E}(\mathbf{x}\mathbf{x}')$. Then, define the **design-independent non-standard target** by

$$\mathbf{x}_0[\hat{M}]'\beta_{\hat{M}}^{(*)} = \mathbf{x}_0[\hat{M}]'\beta[\hat{M}] + \mathbf{x}_0[\hat{M}]' \left(\Sigma[\hat{M}, \hat{M}] \right)^{-1} \Sigma[\hat{M}, \hat{M}^c]\beta[\hat{M}^c],$$

Then, we have for $\mathbf{x}_0 \sim \mathcal{L}$,

$$\mathbb{E} \left(\left[y_0 - \mathbf{x}'_0[\hat{M}]\beta_{\hat{M}}^{(*)} \right]^2 \right) \leq \mathbb{E} \left(\left[y_0 - \mathbf{x}'_0[\hat{M}]\mathbf{v}(\mathbf{Y}) \right]^2 \right),$$

for any function $\mathbf{v}(\mathbf{Y}) \in \mathbb{R}^{|\hat{M}|}$

Asymptotic coverage when p is fixed and $n \rightarrow \infty$

Theorem

Assume that

$$\sqrt{n} [(\mathbf{X}'\mathbf{X}/n) - \Sigma] = O_p(1)$$

and that for any M with $|M| < p$ and for any $\delta > 0$,

$$\sup \left\{ P_{n,\beta,\sigma}(\hat{M} = M | \mathbf{X}) : \beta \in \mathbb{R}^p, \sigma > 0, \|\beta[M^c]\| / \sigma \geq \delta \right\} = o_p(1)$$

Then, for CI obtained by K_1, K_2, K_3, K_4 ,

$$\inf_{\beta \in \mathbb{R}^p, \sigma > 0} P_{n,\beta,\sigma} \left(\mathbf{x}'_0 [\hat{M}] \beta_M^{(*)} \in CI \mid \mathbf{X} \right) \geq (1 - \alpha) + o_p(1)$$

- 1 Introduction and overview
- 2 Confidence intervals for the design-dependent target
- 3 Confidence intervals for the design-independent target
- 4 Simulation study**

Lengths of the confidence intervals

An illustration with the Watershed data set ($p = 9$, $n = 30$)

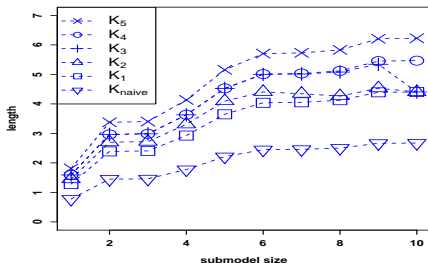


FIGURE: Standardized lengths of the confidence intervals as function of model size. The model of size k is $\{1, \dots, k\}$. $K_5 = \sqrt{p}$ and $K_{naive} = q_{S, n-p, 1-\alpha/2}$ (Student quantile)

Comparison with the confidence intervals of Lee et al. 2016

The confidence intervals of [Lee et al. 2016](#)

- are dedicated to the case where \hat{M} is the LASSO
- are **conditionally valid** (stronger guarantee)

Length comparison between our CIs and those of Lee et al. 2016, on randomly simulated \mathbf{X} (with independent or correlated columns) and \mathbf{Y}

Setting	Lengths	Confidence interval			
		K_1	K_3	K_4	Lee et al.
'Independent'	Median	0.46	0.78	0.78	0.43
	90%-quantile	0.51	0.85	0.85	1.06
'Correlated'	Median	0.56	0.81	0.81	1.42
	90%-quantile	0.90	1.30	1.30	14.3

TABLE: Medians and empirical quantiles of the lengths of the confidence intervals $\bar{C}I$ of Lee et al. 2016 and of those obtained from K_1 , K_3 , and K_4 . The nominal coverage probability is $1 - \alpha = 0.95$, $n = 100$, and $p = 10$.

Minimal coverage probabilities

For $\alpha = 0.05$ and $p = 10$ we evaluate

$$\inf_{\beta \in \mathbb{R}^p, \sigma > 0} P_{n, \beta, \sigma} \left(\mathbf{x}'_0 [\hat{M}] \beta_{\hat{M}}^{(n, \star)} \in CI \mid \mathbf{X} \right),$$

for one realization of \mathbf{X}


Results :

n	model selector	target							
		design-dependent				design-independent			
		$\mathbf{x}'_0 [\hat{M}]' \beta_{\hat{M}}^{(n)}$				$\mathbf{x}'_0 [\hat{M}]' \beta_{\hat{M}}^{(\star)}$			
		K_{naive}	K_1	K_3	K_4	K_{naive}	K_1	K_3	K_4
20	AIC	0.84	0.99	1.00	1.00	0.79	0.97	0.99	0.99
20	BIC	0.84	0.99	1.00	1.00	0.74	0.96	0.98	0.98
20	LASSO	0.90	1.00	1.00	1.00	0.18	0.48	0.61	0.61
100	AIC	0.87	0.99	1.00	1.00	0.88	0.99	1.00	1.00
100	BIC	0.88	0.99	1.00	1.00	0.87	0.99	1.00	1.00
100	LASSO	0.88	0.99	1.00	1.00	0.87	0.99	1.00	1.00

Conclusion

- It is known that it is difficult to construct valid post-model-selection confidence intervals
- Recently, [Berk et al. 2013](#) have proposed confidence intervals for projection-based coefficients
 - ▷ no assumption of correct linear model
 - ▷ valid for all model selection procedure
 - ▷ based on a 'worst-case projection' approach
- We extend the confidence intervals to prediction
 - ▷ exact coverage of the design-dependent target
 - ▷ large p analysis of the length : smaller when small submodels are selected
 - ▷ asymptotic coverage of the design-independent target

The paper :

 **F. Bachoc, H. Leeb, B.M. Pötscher. Valid confidence intervals for post-model-selection predictors,**
<http://arxiv.org/abs/1412.4605>

Thank you for your attention !