Post-model-selection inference

A study on a mixed residual bootstrap procedure

ZHOU Rui

Supervisor: Prof. Stephen M.S. Lee

Abstract The effect of model selection on statistical inference has been widely discussed. The inferential steps ignoring model-selection effect usually result in unreliable conclusions on hypothesis tests and misleading confidence intervals. This report focuses on a mixed residual bootstrap procedure (Lee and Wu, 2018) devised to estimate the distributions of post-model-selection least squares estimators under a linear regression setting. The core of the report is to investigate the empirical performance of the method in the face of different degrees of model misspecification.

Keywords Regression analysis · Model selection effect · Bootstrap · Distribution estimator

1. Introduction

In regression analysis, the conventional statistical inference procedure relies on the knowledge of its true regression model in advance so that some nice properties follow and the validity to perform the conventional statistical tests on the data generated by such a model is granted.

However, in reality the requirement of knowledge on the model in advance is almost impossible to fulfill. Given a set of regression data, it is routine for us to go through the model selection steps. Then there comes a commonly practised yet seriously overlooked problem—the abuse of conventional statistical inference on post-model-selection estimators.

Here is the naive procedure that people usually apply without a second thought. Based on a random sample of regression data as well as a set of user-determined candidate models, a certain data-driven model selection method is employed to suggest one candidate model. Fitting the data to this model gives us some estimates of the parameters. Then it is easy for us to mistakenly believe that it is valid to apply the conventional statistical tests with these estimates. The problem can be confusing unless we realize that the model chosen is random due to the randomness of the data. Probably the next time when we are given another random sample, a different candidate model will be suggested. If we assume the model chosen is known as a priori and perform the traditional statistical tests, it is natural to expect significant results on the regressors that we use the same set of data to pick.

One formal explanation is that the model selection procedure distorts the fixed-model assumption which the conventional statistical tests completely depend on. The form of the sampling distribution of a post-model-selection estimator is no longer guaranteed by any classical statistical theories. Instead, it is determined by the "complex interactions between a suite of possible models and the data to be analyzed", as Berk et al. (2010) mentioned. As a result, the naive procedure where the chosen model is viewed as a priori draws unreliable conclusions on the hypothesis tests and produces some misleading confidence intervals.

In order to take model selection effect into account, we need to consider new inferential procedures. The report focuses on a mixed residual bootstrap procedure proposed by Lee and Wu (2018) under a linear regression setting. It adopts a frequentist approach where a predetermined set of linear regression models not only functions as the set of candidates to choose from, but its intersection closure is also used to generate the residual bootstrap samples. Finally, it provides some estimates for the sampling distributions of the post-model-selection least squares estimators.

Except the ideal case where the true model is included in the set of candidate models, the procedure can be applied to a more common situation where the true model is not included in the candidates or even deviates a lot from a linear model.

Considering the possibility of model misspecification, the report explores how the true model's deviation from linearity may affect its performance. Its empirical performance in the face of different degrees of model misspecification is examined by a few experiments. Conclusions are drawn based on the quality of its estimates for the sampling distributions of the post-model-selection estimators. For comparison, the starting random sample used for bootstrapping additionally goes through the naive procedure which provides unreliable inferential information by assuming the selected model is known as a priori.

2. Mixed residual bootstrap procedure

This section serves as a review of a relatively simplified mixed residual bootstrap procedure proposed by Lee and Wu (2018).

2.1. Problem setting

The problem setting is as follows.

Let us consider a random sample $S = \{(y_1, \boldsymbol{x}_1), (y_2, \boldsymbol{x}_2), \ldots, (y_n, \boldsymbol{x}_n)\}$ modelled by $y_i = \beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i,\ i = 1, 2, \ldots, n,$ where $\{\epsilon_1, \epsilon_2, \ldots, \epsilon_n\}$ and $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ are two independent samples such that $\epsilon_i \sim F_\epsilon$ (a univariate distribution function with zero mean) and $\boldsymbol{x_i} = (x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(p)})^T \sim F_{\boldsymbol{x}}$ (a p-variate distribution function $F_{\boldsymbol{x}}$ with a certain constant p). Our prime interest is the statistical inference of the expected value of y given $\boldsymbol{x} = \boldsymbol{x_0}$. That is, $g(\beta_0, \boldsymbol{\beta}) = \mathbb{E}(y \mid \boldsymbol{x}_0) = \beta_0 + \boldsymbol{x}_0^T \boldsymbol{\beta}$, where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T \in \mathbb{R}^p$.

Given a random sample S, we consider a set of linear regression candidate models $\{M_1, M_2, \ldots, M_q\}$. The set may not include the true model. Each candidate model concerns setting each entry β_i of $\boldsymbol{\beta}$ $, i = 1, 2, \ldots, p$ to zero or not. That is to say, for the j-th candidate model $M_j : y = \beta_0 + \boldsymbol{x}^T \boldsymbol{\beta}^j + \epsilon, j \in \{1, 2, \ldots, q\}$, the value of its coefficient $\boldsymbol{\beta}^j$ is constrained by a certain restriction matrix B_j with dimension $p \times p$ in reduced row echelon form such that $B_j \boldsymbol{\beta}^j = \mathbf{0}$. Then the least squares estimator of $\begin{pmatrix} \beta_0 \\ \boldsymbol{\beta}^j \end{pmatrix}$ subject to the restriction equation $B_j \boldsymbol{\beta}^j = \mathbf{0}$ is given by

$$\begin{pmatrix} \hat{\beta}_0^j \\ \hat{\boldsymbol{\beta}}^j \end{pmatrix} = \left(X^T X\right)^{-\frac{1}{2}} \left\{ I_{p+1} - \left(X^T X\right)^{-\frac{1}{2}} B_j'^{\,T} \left[B_j' \left(X^T X\right)^{-1} B_j'^{\,T} \right]^{-1} B_j' \left(X^T X\right)^{-\frac{1}{2}} \right\} \left(X^T X\right)^{-\frac{1}{2}} X^T Y,$$

where $X = \begin{pmatrix} 1 & \boldsymbol{x}_1^T \\ 1 & \boldsymbol{x}_2^T \\ \vdots & \vdots \\ 1 & \boldsymbol{x}_n^T \end{pmatrix}$, $Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$ and $B_j' = \begin{pmatrix} \mathbf{0} & B_j \end{pmatrix}$.

The choices of the model selector include a wide range of data-driven approaches. The mixed residual bootstrap procedure shall be compatible with any consistent model selector. Two mainstream branches of the model selectors are the information criterion and the sparse estimation which will be further elaborated later. The selection result is a combination of the candidate models $\left(\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_q\right) \in [0, 1]^q$ such that $\sum_{i=1}^q \hat{\lambda}_i = 1$, where $\hat{\lambda}_i$ is associated with candidate model M_i. The deterministic rule to choose one candidate model is a special case where $\left(\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_q\right) \in \{0, 1\}^q$.

Our target is to estimate the sampling distribution $\widetilde{B}$ of the post-model-selection β-statistic

$$\sum_{i=1}^q \hat{\lambda}_i g(\hat{\beta}_0^i, \hat{\boldsymbol{\beta}}^i) - g(\beta_0, \boldsymbol{\beta}).$$

In case of a deterministic model selection, it is also convenient to estimate the sampling distribution $\widetilde{T}$ of the post-model-selection t-statistic

$$\frac{g(\hat{\beta}_0^k, \hat{\boldsymbol{\beta}}^k) - g(\beta_0, \boldsymbol{\beta})}{\sqrt{\widehat{\mathrm{Var}}(g(\hat{\beta}_0^k, \hat{\boldsymbol{\beta}}^k))}},$$

where $(\hat{\beta}_0^k, \hat{\boldsymbol{\beta}}^k)$ is the post-model-selection estimator of $(\beta_0, \boldsymbol{\beta})$ when M_k is chosen, and $\widehat{\mathrm{Var}}(g(\hat{\beta}_0^k, \hat{\boldsymbol{\beta}}^k)) =$

$$\left(1, \boldsymbol{x}_0^T\right)\left(X^T X\right)^{-\frac{1}{2}}\left\{I_{p+1} - \left(X^T X\right)^{-\frac{1}{2}} B_k'^{\,T}\left[B_k'\left(X^T X\right)^{-1} B_k'^{\,T}\right]^{-1} B_k'\left(X^T X\right)^{-\frac{1}{2}}\right\}\left(X^T X\right)^{-\frac{1}{2}}\begin{pmatrix}1\\ \boldsymbol{x}_0\end{pmatrix}\hat{\sigma}^2.$$

2.2. THE PROCEDURE

Then bootstrapping will be done based on each model in the intersection closure $\{M_1, M_2, \ldots, M_{\bar{q}}\}$ of $\{M_1, M_2, \ldots, M_q\}$. The rationale to use such an extended set of candidate models is that by increasing the number of candidate models, we possibly include some better models which are not covered by the original set. However, it needs to be emphasized that the model selection is still among $\{M_1, M_2, \ldots, M_q\}$. The intersection closure is only for the purpose of bootstrap resampling.

The standard residual bootstrap is applied to each model in $\{M_1, M_2, \ldots, M_{\bar{q}}\}$. For a certain model $M_i, i \in \{1, 2, \ldots, \bar{q}\}$, the sample $S = \{(y_1, \boldsymbol{x}_1), (y_2, \boldsymbol{x}_2), \ldots, (y_n, \boldsymbol{x}_n)\}$ is fitted. The corresponding least squares estimator $(\hat{\beta}_0^i, \hat{\boldsymbol{\beta}}^i)$ and the residuals $(\hat{\epsilon}_{i1}, \hat{\epsilon}_{i2}, \ldots, \hat{\epsilon}_{in})$ are obtained. Let $\left(\hat{\epsilon}_{i1}^*, \hat{\epsilon}_{i2}^*, \ldots, \hat{\epsilon}_{in}^*\right)$ be a random sample drawn from the residuals with replacement. A residual bootstrap sample is given by $S_i^* = \{(y_{i1}^*, \boldsymbol{x}_1), (y_{i2}^*, \boldsymbol{x}_2), \ldots, (y_{in}^*, \boldsymbol{x}_n)\}$, where $y_{ij}^* = \hat{\beta}_0^i + \boldsymbol{x}_j^T \hat{\boldsymbol{\beta}}^i + \hat{\epsilon}_{ij}^*$.

We treat S_i^* as if it is the new sample given. Then the model selection procedure can be done for it and the corresponding post-model-selection statistic whose sampling distribution we want to estimate can be calculated. Denote the model selection result based on S_i^* as $\lambda_i^* = \left(\hat{\lambda}_{i1}^*, \hat{\lambda}_{i2}^*, \ldots, \hat{\lambda}_{iq}^*\right)$ and the least squares estimator of fitting S_i^* to each candidate model $k, k = 1, 2, \ldots, q$ as $(\hat{\beta}_{0i}^{k*}, \hat{\boldsymbol{\beta}}_i^{k*})$. The bootstrap post-model-selection β-statistic is given by

$$\sum_{k=1}^{q} \hat{\lambda}_{ik}^* g(\hat{\beta}_{0i}^{k*}, \hat{\boldsymbol{\beta}}_i^{k*}) - g(\hat{\beta}_0^i, \hat{\boldsymbol{\beta}}^i).$$

If we consider the deterministic model selection rule, the bootstrap post-model-selection t-statistic is given by

$$\frac{g(\hat{\beta}_{0i}^{k*}, \hat{\boldsymbol{\beta}}_i^{k*}) - g(\hat{\beta}_0^i, \hat{\boldsymbol{\beta}}^i)}{\sqrt{\widehat{\mathrm{Var}}(g(\hat{\beta}_{0i}^{k*}, \hat{\boldsymbol{\beta}}_i^{k*}))}},$$

where $(\hat{\beta}_{0i}^{k*}, \hat{\boldsymbol{\beta}}_i^{k*})$ is the post-model-selection estimator of $(\hat{\beta}_0, \boldsymbol{\beta})$ given by S_i^* when M_k is chosen.

The procedure illustrated above is based on one bootstrap residual sample. We need to decide the number of bootstrap samples generated by each model in $\{M_1, M_2, \ldots, M_{\bar{q}}\}$ in order to produce a sample of bootstrap post-model-selection β-statistic/ t-statistic. The idea is that more samples should be generated from a model that fits the data better suggested by the sample S. If we have determined the total number of bootstrap samples to generate, then the problem is changed to find the proportion of bootstrap samples generated by each model, which is named as the "mixing weight" and denoted by $\left(w_1, w_2, \ldots, w_{\bar{q}}\right)$, where w_i is associated with M_i. The setting of the mixing weight is relatively intuitive without strict guidelines, but the core requirement is that a higher weight should be assigned to a "better" model. Since there are different kinds of measurement of the models' fitness on the data, it implies a wide range of choices for the mixing weight. Finally, the sample of bootstrap post-model-selection β-statistic/ t-statistic serves as a distribution estimate $\widetilde{B}^* / \widetilde{T}^*$ for $\widetilde{B} / \widetilde{T}$.

3. Simulation examples

The aforementioned mixed residual bootstrap procedure makes no assumption on the unknown true model. Theoretically, it allows situations where the true model is not covered by the linear regression candidate models or even deviates greatly from a linear model. Then, it is natural for us to question:

- Do the residual bootstrap samples generated by the candidate models resemble the true samples? How is the resemblance affected by the degree to which the true model is misspecified by a linear one?
- What is the performance of the bootstrap distribution estimator for a post-model-selection statistic? Can it provide a reliable distribution estimate even in case of severe model misspecification?

In order to answer the above questions, Section 3 probes how the true model's deviation from linearity may affect the reliability of the distribution estimator in the mixed residual bootstrap procedure.

3.1. Problem setting

The data is generated from a true model $y = m(\boldsymbol{x}) + \epsilon$, where

$$m(\boldsymbol{x}; \beta_0, \beta_1, \beta_2, \beta_3, \beta_4) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3\, e^{\beta_4 x_4},$$

$$\boldsymbol{x} \sim N\left(\begin{pmatrix} 2 \\ 0 \\ 1 \\ -1 \end{pmatrix}, \begin{pmatrix} 8.25 & -1.00 & -0.41 & 3.73 \\ -1.00 & 5.85 & 2.74 & 2.34 \\ -0.41 & 2.74 & 4.02 & -1.55 \\ 3.73 & 2.34 & -1.55 & 5.66 \end{pmatrix}\right), \epsilon \sim N(0, 25).$$

Based on a sample $\{(y_1, \boldsymbol{x}_1), (y_2, \boldsymbol{x}_2), \ldots, (y_n, \boldsymbol{x}_n)\}$ of size $n = 200$ generated by the above model, We adopt a certain deterministic rule to select a model from the 16 linear regression models which consider the total combinations of the four regressors. Two kinds of model selectors are considered.

1. Choosing the model M_k with the minimum AIC value, i.e. $\hat{\lambda}_k = 1$, where AIC is calculated as $2d_k + n\ln\frac{\sum_{i=1}^{n}(y_i - \hat{\beta}_0^k - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}^k)^2}{n}$, d_k is the dimension of the model, and $(\hat{\beta}_0^k, \hat{\boldsymbol{\beta}}^k)$ is the least squares estimator.

2. Choose the model with exactly the same regressors eliminated as the result of a LASSO selection, where $\hat{\boldsymbol{\beta}}_{LASSO} = \arg\min_{\boldsymbol{v}\in\mathbb{R}^4}\left\{\frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{v})^2 + a\|\boldsymbol{v}\|_1\right\}$ for some regularization parameter a.

When applying the mixed residual bootstrap procedure, 2000 bootstrap samples in total are generated in one simulation. Two options of mixing weight are tried.

1. BIC method, where $w_k = \frac{e^{-BIC_k}}{\sum_{j=1}^{16} e^{-BIC_j}}$

   BIC value for a certain model M_k is calculated as $d_k \ln n + n\ln\frac{\sum_{i=1}^{n}(y_i - \hat{\beta}_0^k - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}^k)^2}{n}$. Since a smaller value of BIC suggest a model is of better quality, a larger weight is assigned to a model with smaller BIC value.

2. Stationary probability method
   Let each candidate model generate 200 residual bootstrap samples first.
   The stationary probability of the following stochastic matrix serves as the mixing weight.

$$\begin{pmatrix} P_{11} & P_{12} & \cdots & P_{1q} \\ P_{21} & P_{22} & \cdots & P_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ P_{q1} & P_{q2} & \cdots & P_{qq} \end{pmatrix},$$

   where $P_{ij} = \hat{\Pr}(M_j \text{ is selected} \mid \text{ bootstrap samples generated by } M_i)$.

In the experiments, the interest centers on the statistical inference for $\mathbb{E}(y \mid \boldsymbol{x}_0) = m(\boldsymbol{x}_0)$, where $\boldsymbol{x}_0 = (5, 1, -3, 1)^T$. The values of the parameters are unchanged except that of β_4. In each of the following experiments, β_4 is enlarged so that the true model deviates more and more from a linear one.

Experiment No.	$(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)$	Nonlinear index
1	(8, 0.1, -6, 3, 0)	0
2	(8, 0.1, -6, 3, 0.1)	1.9
3	(8, 0.1, -6, 3, 1)	112604.9
4	(8, 0.1, -6, 3, 10)	3.08×10^{71}

where the nonlinear index is to indicate the true model's degree of deviation from a linear model, calculated by $\min_{(\beta_0, \boldsymbol{\beta})} \frac{1}{10000}\sum_{i=1}^{10000}(m(\boldsymbol{x}_i) - \beta_0 - \boldsymbol{x}_i^T\boldsymbol{\beta})^2$, given 10000 random $\boldsymbol{x}$.

3.2. Distribution estimate of post-model-selection t-statistic

Denote the true distribution of the post-model-selection t-statistic as $\widetilde{T}$ and the bootstrap distribution estimate as $\widetilde{T}^*$. Given a specific choice of the model selector and the mixing weight, one simulation in each experiment generates one distribution estimate $\widetilde{T}^*$ for the post-model-selection t-statistic $\frac{\hat{\beta}_0+\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}}-m(\boldsymbol{x}_0)}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_0+\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}})}}$.

In Experiment 1, the true model is $y = 8 + 0.1x_1 - 6x_2 + 3x_3 + \epsilon$, which is covered by the set of candidate models. Figure 3.1 shows the true distribution $\widetilde{T}$, 10 bootstrap distribution estimates $\widetilde{T}^*$ obtained from 10 simulations and the standard normal distribution.

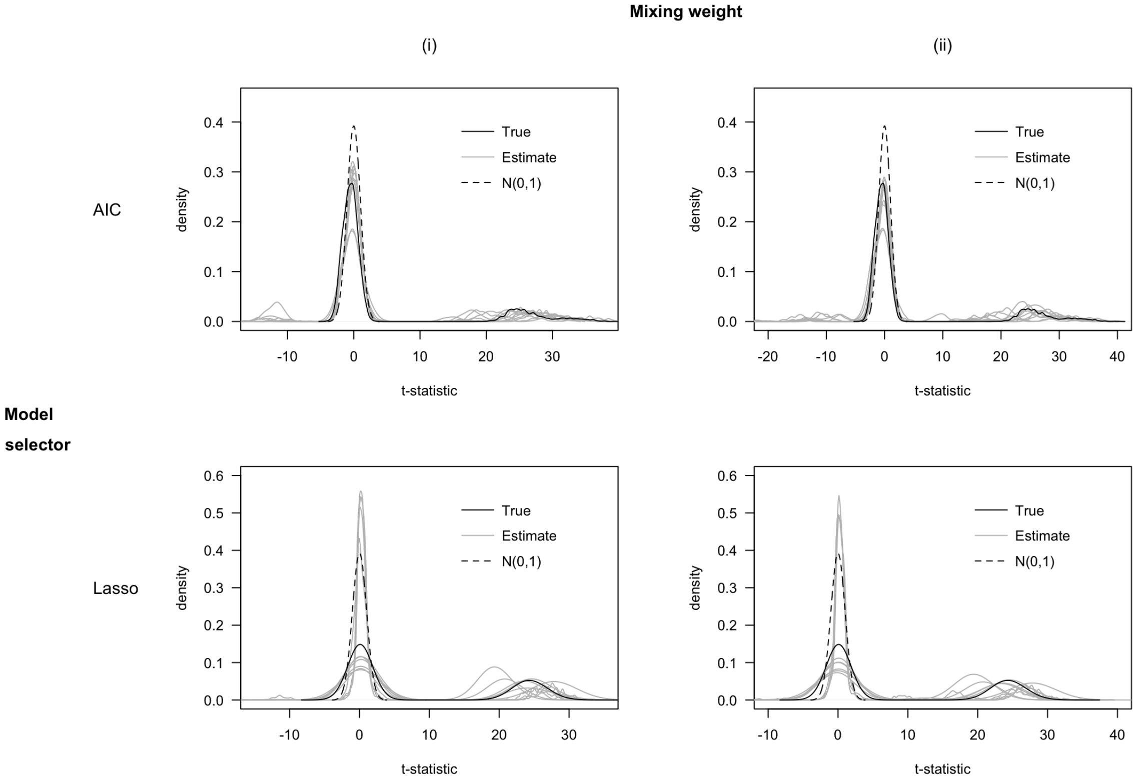

Figure 3.1: Experiment 1: $\widetilde{T}$, $\widetilde{T}^*$ from 10 simulations and $N(0,1)$

There are several pieces of information to tell.

The naive procedure that treats the model chosen as a priori mistakes $\widetilde{T}$ for a standard normal distribution. The graph showing great distinctions between $\widetilde{T}$ and $N(0,1)$ is a direct evidence against the naive procedure and cautions against unwarranted trust in the statistical inference drawn from such a procedure.

Secondly, the bootstrap estimates $\widetilde{T}^*$ generally capture the bi-modal feature of the true distribution in a successful way. It suggests that when the true model is included in the candidate models, the mixed residual bootstrap procedure provides reliable information on the distribution of post-model-selection t-statistic.

Another thing to notice is that it is the choice of model selector affects the true distribution as well as $\widetilde{T}^*$ more. For different mixing weight, as long as the the model selector is fixed, the shape of the true distribution won't have much change.

Experiment 2 (see figure 3.2) starts to add some nonlinearity to the true model, where $y = 8 + 0.1x_1 - 6x_2 + 3x_3\, e^{0.1x_4} + \epsilon$ with nonlinear index 1.9.

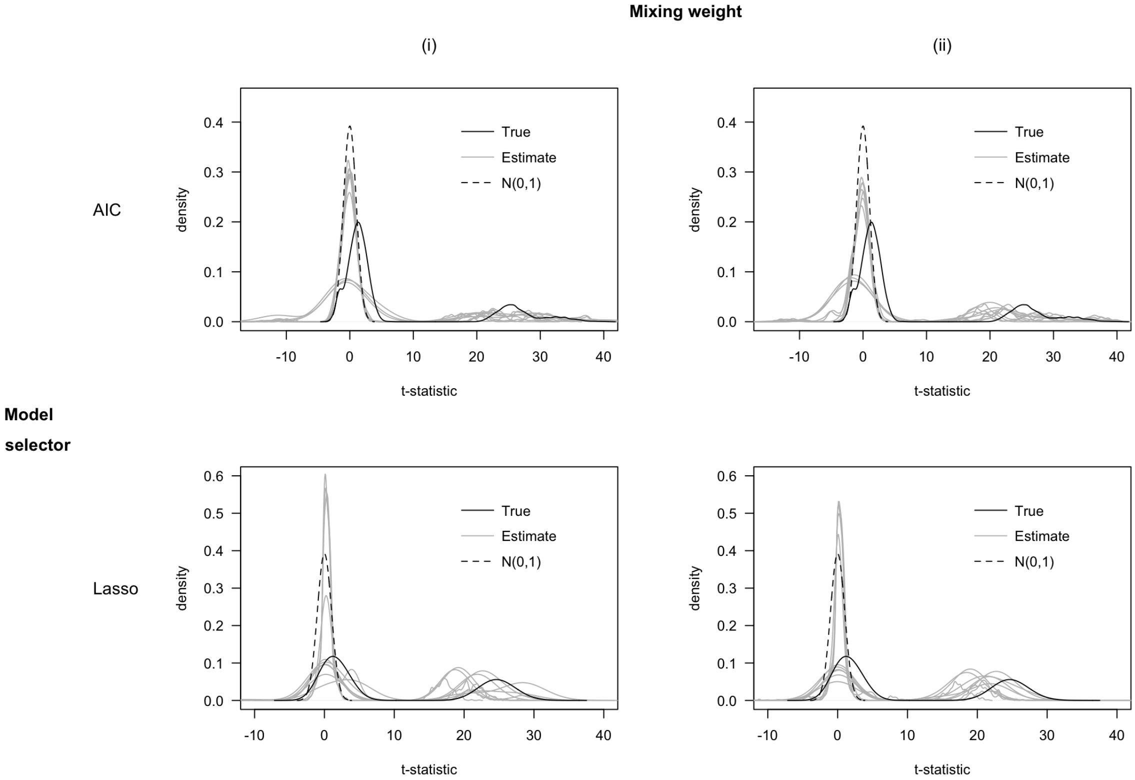

Figure 3.2: Experiment 2: $\widetilde{T}$, $\widetilde{T}^*$ from 10 simulations and $N(0,1)$

The true model is quite similar to a linear one, and probably its patterns can be well captured by some close candidate models so that the bootstrap distribution estimates perform well and their performance is almost as good as those in Experiment 1 when the true model is completely linear.

In Experiment 3 (see figure 3.3 on the next page), by augmenting the value of β_4 to be its 10 times as large, the nonlinearity of the true model is dramatically increased, where $y = 8+0.1x_1-6x_2+3x_3\, e^{x_4}+\epsilon$ with nonlinear index 112604.9. It seems that the mixed residual bootstrap method starts to have difficulty using the linear models to imitate the generating of the true samples. None of the ten estimates shown captures the skewness of the true distribution. However, even though the bootstrap estimates do not coincide with the true distribution that well, its kurtosis is closer to the true distribution than the standard normal distribution is.

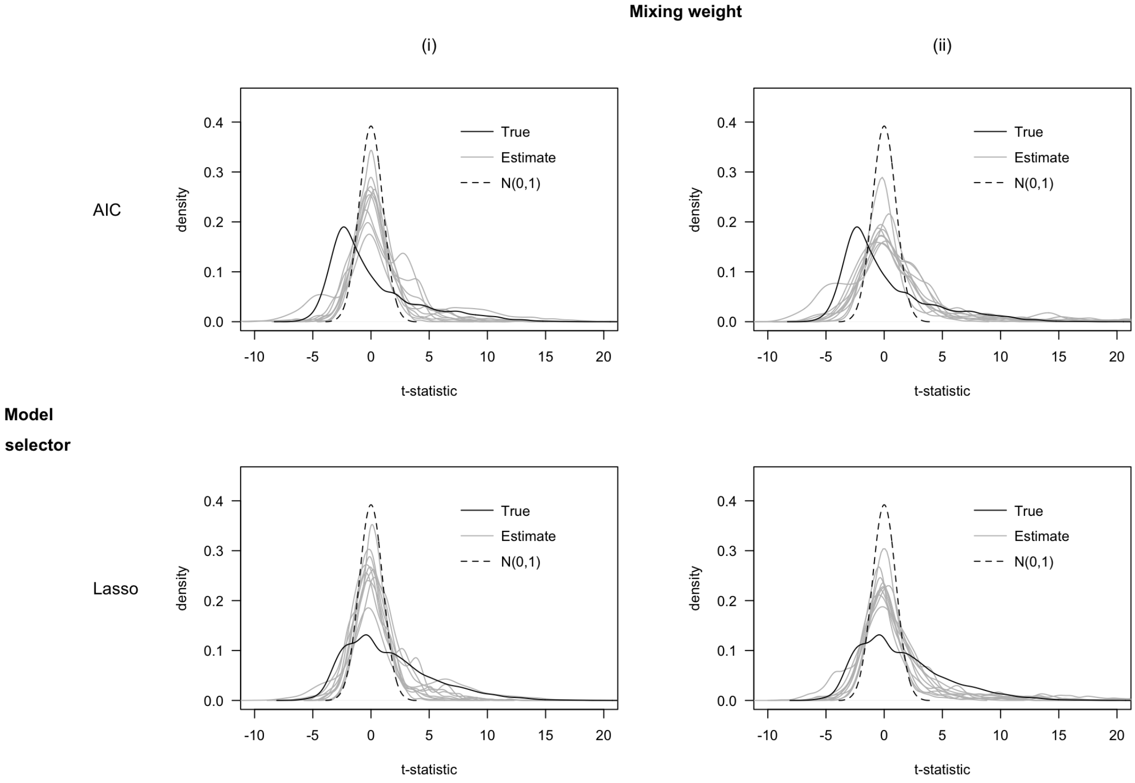

Figure 3.3: Experiment 3: $\widetilde{T}$, $\widetilde{T}^*$ from 10 simulations and $N(0,1)$

Experiment 4 (see figure 3.4 on the next page) is the most extreme case in consideration, where $y = 8 + 0.1x_1 - 6x_2 + 3x_3\, e^{10x_4} + \epsilon$ with exploding nonlinear index 3.08×10^{71}.

Although the performance of the bootstrap estimates is not that well, they are still comparatively better than the standard normal distribution. In this case, the true distribution associated with LASSO is bi-modal, while the true distribution associated with AIC is triple-modal. But both of them put fewer density around zero. The mixed residual bootstrap procedure works effectively in terms of pressing the density of a normal distribution down at the place around zero. Also, it can be clearly seen that the bootstrap estimates try to capture the triple-modal feature when the model selector is AIC.

The four graphs (figure 3.1–3.4) on the distribution estimates are arranged in a way that the degree of model misspecification is increasing. When the model is more and more severely misspecified, it is also increasingly difficult for the distribution estimator to capture the shape of the true distribution. However, it is worth noticing that compared with the standard normal distribution, the bootstrap estimates unexceptionally perform better in terms of approximating the true distribution in the four scenarios.

However, we should note that even though the distribution estimates are closer to the true distribution in shape, it does not a-hundred-percent ensure that the confidence intervals generated by the distribution estimates would have higher coverage probability than those generated by the standard normal distribution, which is indeed true as the next subsection discusses.

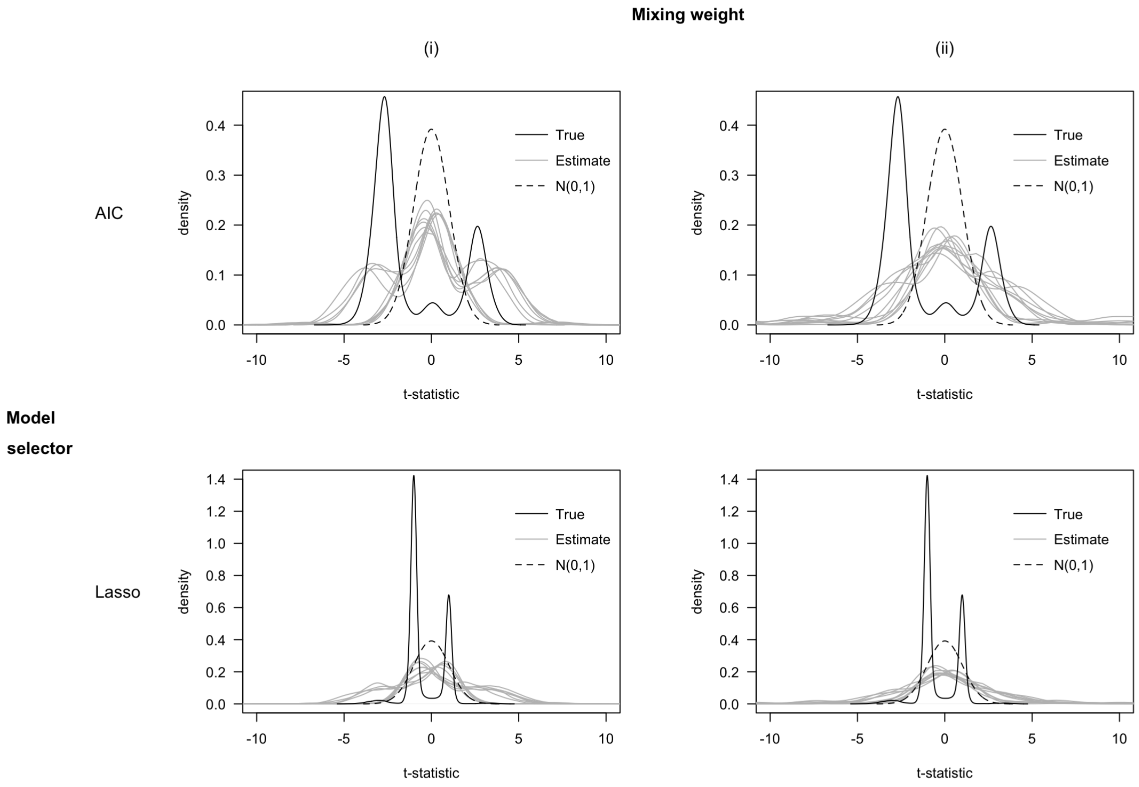

Figure 3.4: Experiment 4: $\widetilde{T}$, $\widetilde{T}^*$ from 10 simulations and $N(0,1)$

3.3. Coverage probability of confidence intervals

After obtaining the bootstrap distribution estimates $\widetilde{T}^*$ and $\widetilde{B}^*$ for

$$\frac{\hat{\beta}_0 + \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}} - m(\boldsymbol{x}_0)}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_0 + \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}})}} \text{ and } \hat{\beta}_0 + \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}} - m(\boldsymbol{x}_0) \text{ respectively,}$$

the 95% confidence intervals can be calculated as follows.

1. 95% Bootstrap t confidence interval

$$\left[\hat{\beta}_0 + \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}} - \widetilde{T}^*_{0.975}\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_0 + \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}})}, \hat{\beta}_0 + \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}} - \widetilde{T}^*_{0.025}\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_0 + \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}})}\right]$$

2. 95% Bootstrap quantile confidence interval

$$\left[\hat{\beta}_0 + \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}} - \widetilde{B}^*_{0.975}, \hat{\beta}_0 + \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}} - \widetilde{B}^*_{0.025}\right]$$

3. 95% Naive confidence interval

$$\left[\hat{\beta}_0 + \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}} - 1.96\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_0 + \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}})}, \hat{\beta}_0 + \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}} + 1.96\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_0 + \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}})}\right]$$

   This is the confidence interval obtained by the naive procedure where the post-model-selection t-statistic is thought to follow a standard normal distribution.

We shall compare the performance of the above confidence intervals with the following true 95% confidence intervals constructed by bootstrap t and bootstrap quantile methods respectively

$$\left[\hat{\beta}_0 + \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}} - \widetilde{T}_{0.975}\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_0 + \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}})}, \hat{\beta}_0 + \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}} - \widetilde{T}_{0.025}\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_0 + \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}})}\right], \text{and}$$

$$\left[\hat{\beta}_0 + \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}} - \widetilde{B}_{0.975}, \hat{\beta}_0 + \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}} - \widetilde{B}_{0.025}\right].$$

Then let us investigate the quality of these confidence intervals by their coverage probabilities. For a certain confidence interval, the coverage probability is estimated by averaging the counts of the event of $m(\boldsymbol{x}_0) \in$ the confidence interval over 1000 simulations. That is,

$$\frac{1}{1000}\sum_{i=1}^{1000} \mathbf{1}\{m(\boldsymbol{x}_0) \in \text{ confidence interval of the i-th simulation}\}$$

In figure 3.5 on the next page, the sold black and red lines correspond to the estimated coverage probabilities of the exact 95% confidence intervals. They slightly fluctuates around the 95% level due to some randomness ($\pm 1.96\sqrt{\frac{0.95\times 0.05}{1000}} \approx \pm 0.0135$, theoretically). The dashed black and red lines correspond to the 95% bootstrap quantile CI and the bootstrap t CI. The grey line refers to the 95% naive confidence interval.

Generally speaking, the coverage probabilities of the bootstrap confidence intervals are closer to the "alleged" 95% confidence level, compared to those of the naive confidence intervals.

The bootstrap confidence intervals usually perform well in case of a slight degree of model misspecification, which is consistent with our impression on its nice distribution estimates in Experiment 1 and 2.

When the model is severely misspecified, as the results of Experiment 3 and 4 show, the bootstrap distribution estimates tend to produce conservative confidence intervals whose coverage probabilities are larger than the said confidence level. Experiment 4 presents such an example. When LASSO is employed to choose model, the estimated coverage probability of the bootstrap quantile confidence interval reaches 99.7% and 99.8% (see table A.4) respectively under the two choice of mixing weights.

It is also interesting to learn that the coverage probability of the naive confidence interval happens to have a good performance with a coverage probability of 96.2% (table A.4) in Experiment 4 using LASSO. A more funny thing is that such a performance is in stark contrast with the performance of the naive CI on the same data but using AIC as the model selector

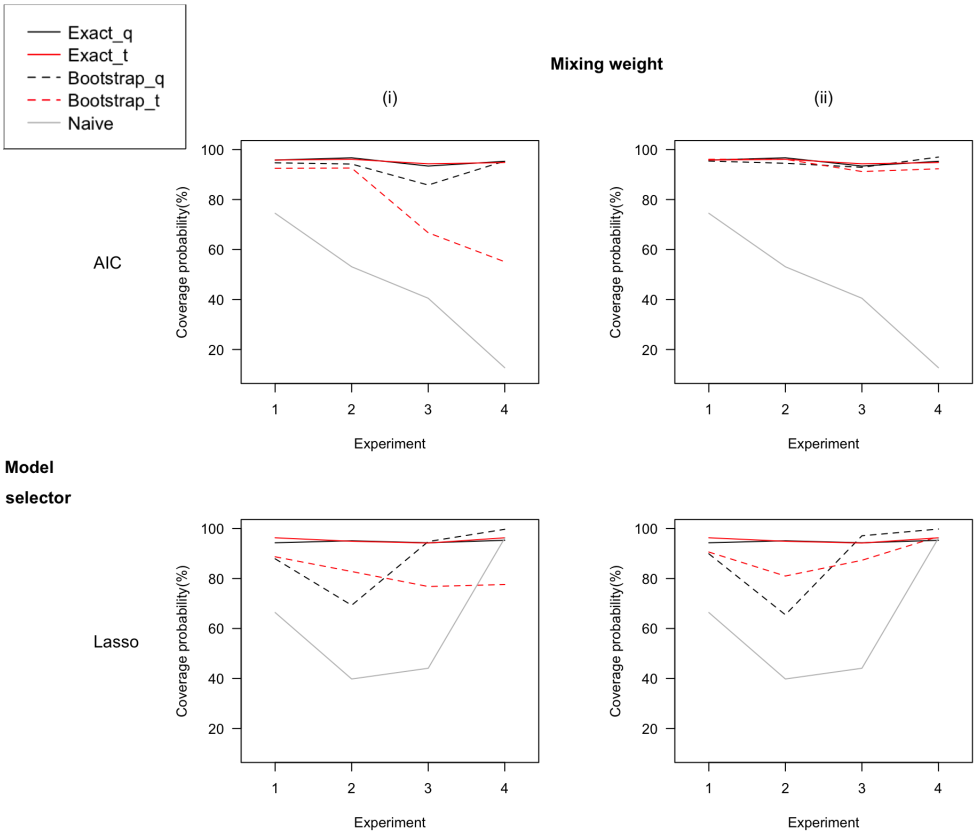

Figure 3.5: Coverage probabilities of the 95% confidence intervals (Detailed values can be found in Appendix)

instead where its coverage probability drastically drops to only 12.7% (table A.4). This phenomenon suggests that for a naive confidence interval, its performance is extremely unstable and there is no guarantee on its quality. When dealing with the same data, a different choice of model selector may completely change its coverage probability.

3.4. Experiment 5

Now let us turn back to the model in Experiment 4, which is $y = 8 + 0.1x_1 - 6x_2 + 3x_3\, e^{10x_4} + \epsilon$ with exploding nonlinear index 3.08×10^{71}.

It is a convention for people to apply log-transformation on the data when there are signs of exponential patterns. That is equivalent to change the model in Experiment 4 to

$$y = \operatorname{sgn}\left(8 + 0.1x_1 - 6x_2 + 3x_3\, e^{10x_4} + \epsilon\right) \log\left(8 + 0.1x_1 - 6x_2 + 3x_3\, e^{10x_4} + \epsilon\right),$$

where $\operatorname{sgn}(\cdot)$ is the function taking the sign of its argument.

This is also a case of model misspecification since such a true model is not covered by the linear regression models. Its nonlinear index is 138.6 whose nonlinearity is between the models in Example 2 and 3. Then let us have a look at the distribution graph as well as the coverage probability table.

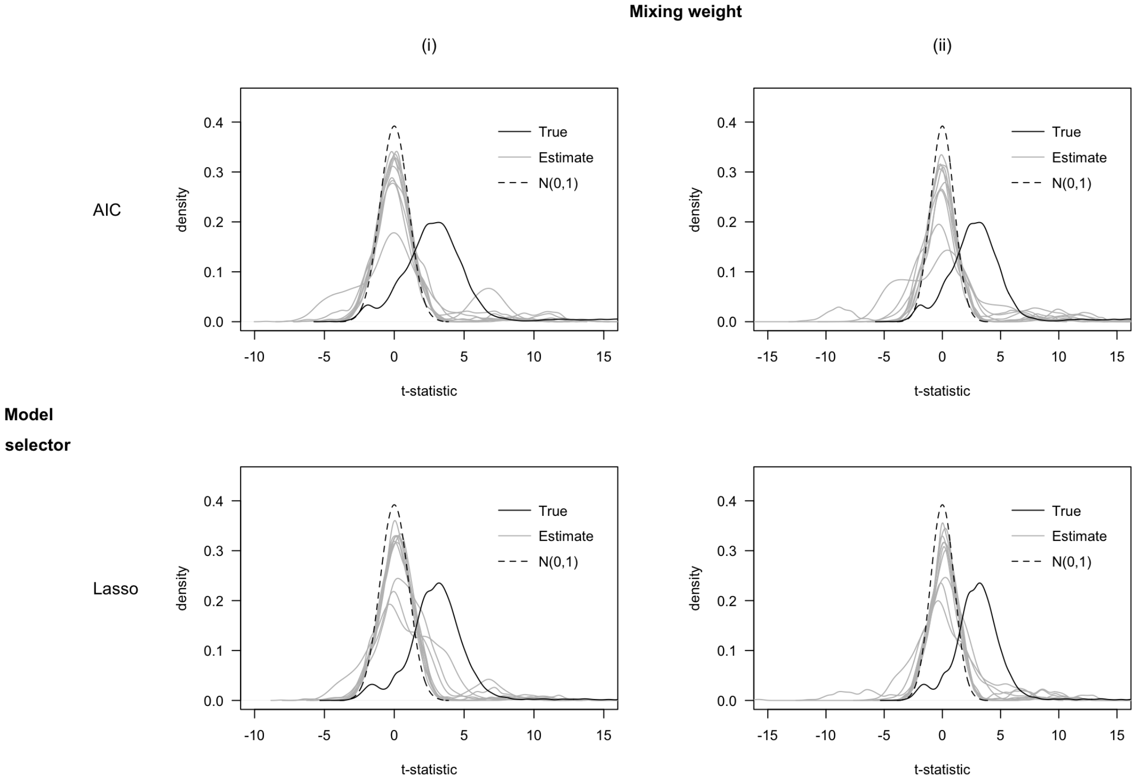

Figure 3.6: Experiment 5: $\widetilde{T}$, $\widetilde{T}^*$ from 10 simulations and $N(0,1)$

95% CI	Model selector	Exact	Mixing (i)	weight (ii)	Naive CI
Bootstrap quantile CI	AIC LASSO	95.2 96	84.6 96.8	94.8 96.9	
Bootstrap t CI	AIC LASSO	94.9 95.1	88.3 79.3	92.4 82.6	29.2 25.2

Table 3.1: Experiment 5: Coverage probabilities (×100) of 95% confidence intervals constructed by different methods

The true distribution behaves like a log-normal distribution with nonzero mean, while the bootstrap estimates approximately centre on zero. Comparing the coverage probabilities of the bootstrap confidence intervals and the naive confidence intervals, we are further convinced that in case of mild model misspecification the mixed residual bootstrap procedure usually has a good performance in generating reliable confidence intervals. In addition, the result emphasizes the false statistical power resulted from a naive procedure neglecting the model selection effect.

4. Conclusion

Post-model-selection inference is tricky. On the one hand, it is common for people to neglect the model selection effect in statistical inference and perform the classical statistical tests as if the selected model is known as a prior. One the other hand, the flawed statistical inference after model selection may lead to serious mistakes in decision making, especially terrible for subjects like criminology.

However, even though people realize the pitfalls in post-model-selection inference, there is no neat way to take the model selection effect into account. What's more, Leeb and Pötscher (2008) proved that no estimator for the distribution of a post-model-selection estimator is uniformly consistent.

Among the existing methods to compensate for the loss of statistical power after model selection, the report studies a mixed residual bootstrap procedure. Some insights are provided based on its empirical performance in case of different degrees of model misspecification. After several experiments that increase the nonlinearity of the true model gradually, the robustness of the method is demonstrated. It usually provides some relatively reliable inference when the unknown model is approximately linear. However, in case of severe model misspecification, the inference based on the distribution estimator tend to be more conservative than the alleged significance level.

Acknowledgement

I would like to express my sincere gratitude to Prof. Stephen LEE for his time, patience and support. I greatly benefit from his heuristic teaching and academic insights. I also would like to thank Mr. Lawrence TSUI for his IT support. I could not have finished the experiments in time without his help.

References

Berk, R., Brown, L., and Zhao, L. (2010). Statistical inference after model selection. *Journal of Quantitative Criminology*, 26(2):217–236.

Lee, S. M. S. and Wu, Y. (2018). A bootstrap recipe for post-model-selection inference under linear regression models. *Biometrika*, 105(4):873–890.

Leeb, H. and Pötscher, B. M. (2008). Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory*, 24(2):338–376.

A. Tables of coverage probabilities

95% CI	Model selector	Exact	Mixing	weight	Naive
			(i)	(ii)	CI
Bootstrap	AIC	95.8	94.7	95.4	
quantile CI	LASSO	94.3	87.9	89.9	
Bootstrap	AIC	95.4	92.5	96.1	74.5
t CI	LASSO	96.3	88.7	90.6	66.4

Table A.1: Experiment 1: Coverage probabilities (×100) of 95% confidence intervals

95% CI	Model selector	Exact	Mixing	weight	Naive
			(i)	(ii)	CI
Bootstrap	AIC	96.7	94.2	94.5	
quantile CI	LASSO	95.1	69.3	65.5	
Bootstrap	AIC	96.1	92.6	96.1	51.3
t CI	LASSO	94.9	82.8	81	39.8

Table A.2: Experiment 2: Coverage probabilities (×100) of 95% confidence intervals

95% CI	Model selector	Exact	Mixing	weight	Naive
			(i)	(ii)	CI
Bootstrap	AIC	93.4	85.8	92.9	
quantile CI	LASSO	94.3	94.8	97.1	
Bootstrap	AIC	94.3	66.7	91.2	40.5
t CI	LASSO	94.2	76.8	87.3	44.1

Table A.3: Experiment 3: Coverage probabilities (×100) of 95% confidence intervals

95% CI	Model selector	Exact	Mixing	weight	Naive
			(i)	(ii)	CI
Bootstrap	AIC	95.3	95.4	97	
quantile CI	LASSO	95.3	99.7	99.8	
Bootstrap	AIC	94.8	55.1	92.3	12.7
t CI	LASSO	96.3	77.6	96.8	96.2

Table A.4: Experiment 4: Coverage probabilities (×100) of 95% confidence intervals