Spring 5-2017

On Post-selection Confidence Intervals in Linear Regression

Xinwei Zhang
Washington University in St. Louis

Washington University in St. Louis

Department of Mathematics

On Post-selection Confidence Intervals in Linear Regression

by

Xinwei Zhang

A thesis presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Master of Arts

May 2017
Saint Louis, Missouri

Contents

List of Tables

List of Figures

Acknowledgments

I am deeply grateful for the guidance of my advisor and committee members throughout the process of writing this thesis. I would like to thank my advisor, Dr. Todd Kuffner, for his willingness to work with me even before I finished all the course requirements for a Master's degree. I greatly appreciate the freedom and support he provided me in searching a thesis topic. And I thank him for the great amount of time he invested in working with me and valuable insight he provided into my work. I am also thankful for his patience and encouragement every time I encountered difficulties and even felt depressed in writing this thesis. I could not have had a better mentor than Dr. Kuffner. I also would like to thank my committee members, Dr. Kuffner and Dr. Figueroa-López, for their knowledgeable advice, time and effort.

Last but certainly not least, I want to thank my parents and all my family members. Without their support, I could never imagine the achievements I have made so far.

Xinwei Zhang

Washington University in Saint Louis
May 2017

ABSTRACT OF THE THESIS

On Post-selection Confidence Intervals in Linear Regression

by

Xinwei Zhang

Master of Art in Statistics

Washington University in St. Louis, May 2017

Research Advisor: Professor Todd Kuffner

The general goal of this thesis is to investigate and examine some issues about post-selection inference which arises from the setting where statistical inference is carried out after a data-driven model selection step. In this setting, the classical inference theory which requires a fixed priori model becomes invalid since the selected model is a result of random event. Hence, a common practice in applied research which ignores the model selection and builds up confidence interval will result in misleading or even false conclusion. In this thesis, specifically, we first discusses some examples to show how the classical inference theory loses validity after selection. Then we focus on the scenario of linear regression, and review two different interpretation views of parameters, i.e., full model view and and submodel view. We study the simultaneous post-selection inference solution under submodel view provided by Berk et al. [*Ann. Stat.* **41** (2013) 802-837] and carry out simulation to examine the results of Leeb, Pötscher and Ewald. [*Stat. Sci.* **30** (2015) 216-227].

Chapter 1

Introduction

Within the framework of classical statistical theory, the model which generates data is assumed to be known. After obtaining data from a priori model, we can build up a valid statistical test or confidence interval to examine properties of the parameters of the priori model. However, in many statistical practices, a priori model is rarely presumed before exploring the data. More commonly, data analysts use sophisticated tools to search through a large pool of candidate models and then report inferential conclusions based on the selected model. Taylor and Tibshirani (2015) describes this as an "industrialization of statistical methods" in response to the technological advance in science and industry.

Typically, analysts start with a collection of competing models based on a given dataset. The collection might be all submodels within the full model (using all variables) or some submodels obtained through certain restriction on the full model. The first step is to apply a model selection procedure to choose the most parsimonious model. The selection procedure can be based on a hypothesis test, on the optimization of a penalized goodness-of-fit criterion, final prediction error, or cross-validation. After selection, the second step is to estimate underlying model parameters from the most parsimonious model for statistical inference. The estimators obtained after model selection are called "post-model-selection estimators"

in Leeb and Pötscher (2005). For consistency with the literature, we also use this name to call estimators resulting from model-selection in this thesis.

However, this practice is always problematic in the inference part, since analysts would ignore the model selection effect and use the nominal distribution of post-model-selection estimator to make inference. For example, in the case of linear regression where error is assumed to be independently identically normally distributed with known variance, after model selection, analysts still treat the distribution of least squares estimator as normal distribution and construct $(1-\alpha)$ two-sided confidence interval through $\alpha/2$ and $1-\alpha/2$ quantile of nominal normal distribution. It turns out that, because of the stochastic nature of data-driven model selection procedure, the sampling distribution of the post-model-selection least squares estimator is no longer the same as nominal normal distribution. Hence, the inferential conclusions will be misleading.

There is a lot of evidence in the literature which reports similar detrimental impacts on the subsequent inference after model selection, for example. In particular, Leeb and Pötscher (2005) shows that, even using consistent model selection methods (in the sense that it asymptotically select the true model with probability 1), the sampling distribution of post-model-selection estimator will not be the same as nominal distribution asymptotically. It falsifies the argument that a consistent model selection procedure would allow one to use the standard asymptotic result which applies when a model is assumed a priori.

These broadly-existing problems in statistical inference after model selection have now drawn considerable attention from statisticians, and are promoted as an active research area called post-selection inference. In the recent article, Berk, Brown, Buja, Zhang, and Zhao (2013) proposed a new class of confidence intervals called PoSI-intervals. The PoSI-intervals reduce

the post-selection inference problem in linear model to one simultaneous inference which consider coefficient estimates in all submodels.

An important tenet of the PoSI procedure which yields inferential validity under all possible models considered by the selection procedure, is an interpretation of the linear model parameters known as the submodel viewpoint, i.e. viewing deselected parameters as nonexistent, in contrast to full model interpretation, i.e. viewing deselected parameters as estimating coefficient to be zero. Following the submodel view, Berk et al. (2013) proposes to consider confidence intervals for an unconventional quantity of interest which depends on the submodel, while the conventional quantity of interest is a fixed parameter of a data-generating model. Then they provide a valid solution to the problem of post-selection inference on the new covering target. In this thesis, we review the two different interpretations with their theoretical consequences, and examine how PoSI-intervals achieve desired minimal-coverage probability after model selection.

We also follow the idea in Leeb, Pötscher, and Ewald (2015) to carry out simulations and examine the performance of "naive" confidence intervals which are constructed regardless of model selection, and PoSI intervals. We do not only consider the unconventional coverage target in Berk et al. (2013), but also examine the coverage performance for the conventional target.

The outline for the thesis is as follows. In Section 2, we show some examples to illustrate the problem of post-selection inference. In Section 3, we first show two different interpretations, full model view and submodel view, basic assumptions and coverage target under submodel view. In Section 4, we review the universally valid post-selection inference in providing confidence guarantees of PoSI-intervals as proposed in Berk et al. (2013). In Section 5, we report reylsts of a simulation study in which we compare these different types of confidence

intervals including "naive", PoSI, PoSI1 and Scheffé confidence intervals based on their empirical minimal coverage probabilities.

Chapter 2

Examples of Post-selection Inference

In this section, we provide some examples to illustrate the problem of invalidation of classical inference after model selection. The nature of this invalidation is that the data-driven model selection is stochastic but is not accounted for in classical theory. Thus, the sampling distribution is no longer the same after model selection. Drawing ideas from Fithian, Sun, and Taylor (2014) and Benjamini and Yekutieli (2005), we provide a first example to show some intuition.

Example 1 (File Drawer Effect) Suppose we collect n independent observations, each of which follows $Y_i \sim N(\mu_i, 1)$. We focus only on the apparently large effects, selecting only the indices i for which $|Y_i| > 1$. We denote the selected index set as $I = \{i : |Y_i| > 1\}$.

Part I: First, we wish to test $H_{0,i} : \mu_i = 0$ for each $i \in I$ at the $\alpha = 0.05$ significance level. Then if we ignoring the selection effect, the classical test that rejects $H_{0,i}$ when $|Y_i| > 1.96$ is invalidated by the selection.

Now, what exactly is invalid in this test? The answer is that, among the selected effect, the fraction of false rejections will not be 0.95 anymore. Let n_0 be the number of true null effects and suppose $n_0 \to \infty$ as $n \to \infty$. Then, as $n \to \infty$, the fraction of false rejections among

the true nulls we select is

$$\begin{aligned}\frac{\text{number of false rejections}}{\text{number of selected true nulls}} &= \frac{\sum_{i:H_0 \text{ is true}} \mathbb{1}_{i\in I \text{ and } H_{0,i}\text{is rejected}}}{\sum_{i:H_0 \text{ is true}} \mathbb{1}_{i\in I}} \\ &\longrightarrow \frac{n_0 \cdot \mathbb{P}_{H_{0,i}}\left[i \in I \text{ and } H_{0,i} \text{ is rejected}\right]}{n_0 \cdot \mathbb{P}_{H_{0,i}}[i \in I]} \\ &= \mathbb{P}_{H_{0,i}}[H_{0,i} \text{ is rejected } | i \in I] \\ &= \frac{\Phi(-1.96)}{\Phi(-1)} \approx 0.16.\end{aligned}$$

Part II: Now, suppose we wish to further construct confidence interval for selected observations. Under classical theory, we construct $1-\alpha$ confidence interval in the "naive" way which is $Y_i \pm Z_{1-0.5/2}$, where $Z_{1-\alpha}$ is $1-\alpha$ quantile of standard normal distribution. (In this thesis, we abuse the notation $a \pm b$ to denote interval $[a-b, a+b]$)

Then the nominal $1-\alpha$ CI is invalid in the sense that conditional coverage rate may not be $1-\alpha$. For simplification, suppose that $\mu_i \equiv \mu$ fixed, then the conditional coverage probability—the number of times that the true parameter is covered by the CI divided by the number of times that the observation is selected—can be expressed as

$$\begin{aligned}\frac{\text{number of covering CIs}}{\text{number of selected observations}} &\longrightarrow \frac{\mathbb{P}_\mu\left[\mu \in Y_i \pm Z_{1-0.5/2} \text{ and } |Y_i| > 1\right]}{\mathbb{P}_\mu[|Y_i| > 1]} \\ &= \mathbb{P}_\mu\left[|Y_i - \mu| < 1.96 \middle| |Y_i| > 1\right].\end{aligned}$$

The conditional coverage probability depends on the true value of μ. For five different values of μ: 0, 0.5, 1, 2, and 4, the conditional coverage probabilities are, respectively, 0.84, 0.87, 0.91, 0.97, 0.95.

As we can see from the **Example 1**, the conditional false rejection probability and conditional coverage probability are no longer $1-\alpha$ as desired. Most researchers would agree that this is a sign of invalidation of inference. However, one may argue that the inference may still be valid in the view of unconditional probability of a false positive error. For example, in part I, the unconditional false rejection probability will still be less than α. This unconditional false rejection probability actually equals the False Discovery Rate (FDR) in Benjamini and Hochberg (1995). And, in part II, the unconditional probability of constructing a non-covering CI is still be controlled to be less than α, since

$$\mathbb{P}\left[\mu \notin CI, CI \text{ constructed}\right] \leq \mathbb{P}\left[\mu \notin CI\right] \leq \alpha.$$

This also equals to the False Coverage-statement Rate (FCR) in Benjamini and Yekutieli (2005). The FCRs for the above five parameter values of μ are, respectively, 0.05, 0.05, 0.05, 0.03, 0.05.

The conditional probability of a false positive error and unconditional probability of a false positive error appear to conflict with one another, when both viewed as performance measures for post-selection inference. However, we need to note that, after selection, the unconditional probabilities of a false positive error is no longer $\mathbb{P}_{H_0}\left[H_0 \text{ is rejected}\right]$ when making hypothesis test and $\mathbb{P}\left[\mu \notin CI\right]$ when constructing CI, since either $\{H_0 \text{ is rejected}\}$ or $\{\mu \notin CI\}$ is not a well-defined event as we only make inference for selected observations. Therefore, controlling

unconditional probability of a false positive error after selection does not have the same meaning as it has without selection.

In addition, we want to point out a big problem if we validate the unconditional probability of a false positive error or its extensions like FDR and FCR as performance measures of post-selection statistical inference. These methods generally treats no selections as making zero error, which implicitly take no selections as making 100% correct. Take FCR as an example. A procedure which never reports a CI will have a perfect FCR of 0. And a procedure which rarely reports a CI will also have a good FCR performance, even if it has a high error rate conditioning on the event that it reports a CI.

Therefore, we believe that even the idea of controlling unconditional false error is not fully justified and in fact is theoretically controversial. In above cases, although the unconditional probabilities of a false positive error is under control, it dose not validate the classical theory in the post-selection inference. Furthermore, we can show that, even using the unconditional criterion, the problem can not be solved automatically.

Example 2 (Linear Model Selection with Fixed Parameter) Drawing ideas from both Berk, Brown, and Zhao (2010) and Leeb and Pötscher (2005), we consider an example of linear regression.

Consider now there is a true underlying model,

$$y_i = \alpha_1 X_{1i} + \alpha_2 X_{2i} + \delta_i \tag{2.1}$$

where the $\delta_1 \sim N(0, \sigma^2)$. Researchers don't know what the true coefficients are, and decide to estimate the parameters by linear regression.

For our purpose, we mainly want to illustrate the influence of model selection on the unconditional error control. The unconditional error is not well-defined in the **Example 1**. To avoid the problem, we consider the situation where the parameter of interest is α_1, and X_1 is always fixed in the pool of candidate models for selection. Hence consider choosing from two candidate models and making inference on parameter α_1. Since $\boldsymbol{X}_1$ is always included during model selection, the probability of any event conditional on $\boldsymbol{X}_1$ being selected is the same as its marginal probability. Therefor, the unconditional error for inference on α_1 should be well-defined. Call two candidate models as $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$, i.e., $\boldsymbol{M}_1$ is

$$y_i = \beta_1 X_{1i} + \varepsilon_i \tag{2.2}$$

and $\boldsymbol{M}_2$ is

$$y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i. \tag{2.3}$$

Without loss of generality, we use some simulation results for illustration. For analytical results, we refer to Leeb and Pötscher (2005). For this simulation, we first draw predictors from a multivariate normal distribution with mean and the variance and covariance specified as: $\mathbf{E}[X_1] = 8$, $\mathbf{E}[X_2] = 1$, $\sigma^2_{X_1} = 20$, $\sigma^2_{X_2} = 15$, $cov(X_1, X_2) = 14$. The sample size is 20. The setting is of no importance in the results, and is not designed to illustrate the worst case scenario. The predictors are fixed in the simulation. Then we draw 10000 Monte Carlo realizations $\boldsymbol{y}^{(n)} = (y_1^{(n)}, \cdots, y_{20}^{(n)})^T$ from the true model (2.1) where the variance of error δ is $\sigma^2_\delta = 15$. For each Monte Carlo sample $\boldsymbol{y}^{(n)}$, the AIC is used to select from the two candidates model $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$.

Case 1: Consider the case that $\alpha_1 = 1$, $\alpha_2 = 2$ and the variance $\sigma^2_\varepsilon = \sigma^2_\delta = 15$ is known a priori. In this case, $\boldsymbol{M}_2$ is considered to be the true model. We plot the density of estimated

regression coefficients $\hat{\beta}_1$. The black solid line represents the density when $\boldsymbol{M}_2$ is known to be correct a priori and red dashed line represents the post-selection sampling distribution. Since the variance of the error is known, if we know $\boldsymbol{M}_2$ is true a priori, then by classical theory, we have $\hat{\beta}_1$ follows a normal distribution. After incorporating the model selection, we have 5121 times $\boldsymbol{M}_2$ selected and 4879 times $\boldsymbol{M}_1$ selected. The probabilities for selecting $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ almost equal.

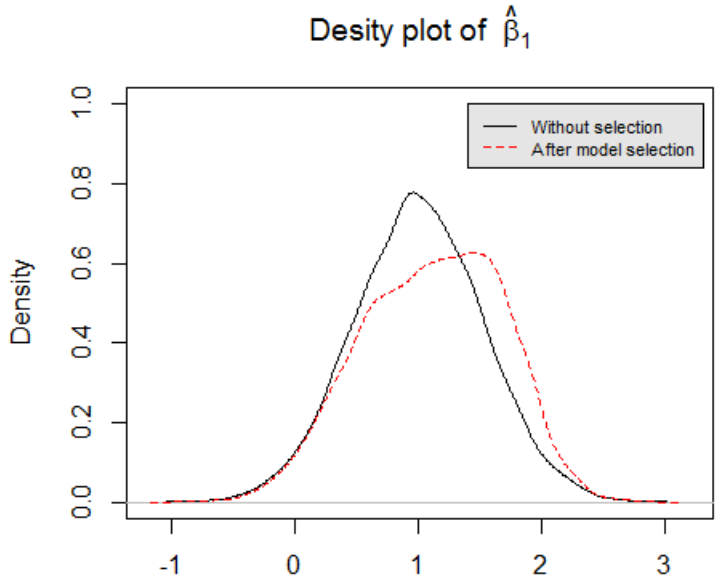

Figure 2.1: The density plot of $\hat{\beta}_1$ (The solid black line is the density where $\boldsymbol{M}_2$ is assumed to be correct a priori and the red dashed line is the post-selection density of $\hat{\beta}_1$)

Now, we want to check that after model selection, the distribution of $\hat{\beta}_1$ still has the same shape. The results are shown in Figure 2.1. As we can see, after model selection, the density of β_1 is no longer normally distributed. This implies that a confidence interval derived from

a z-statistics will not provide valid unconditional error control. However, the situation here is not that bad because the difference is small.

Case 2: Consider the case that $\alpha_1 = 1$, $\alpha_2 = 2$ and the variance σ_ε^2 is not known and must be estimated. We plot the density of t-values $t = \hat{\beta}_1/\hat{\sigma}_{\beta_1}$ instead of the estimated regression coefficients $\hat{\beta}_1$. The black solid line and red dashed line represent the unselected density and post-selection density respectively. The distribution of t-values is more informative because it take both regression coefficients and their standard errors into account. And you can induce the performance of confidence intervals based on the t-statistics. In this case, $\boldsymbol{M}_2$ is still considered to be the true model. Since the variance of error is unknown, if we know $\boldsymbol{M}_2$ is true a priori, then by classical theory, we have $\hat{\beta}_1/\hat{\sigma}_{\beta_1}$ follows a t-distribution. This time, after incorporating the model selection, we have 5731 times $\boldsymbol{M}_2$ selected and 4369 times to select $\boldsymbol{M}_1$ selected out of 10000 experiments in total.

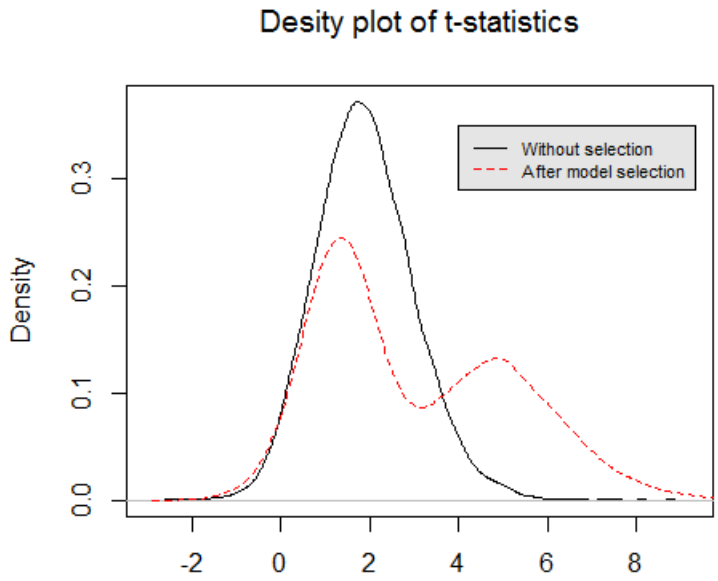

Figure 2.2: The density plot of t-statistics (The solid black line is the density where $\boldsymbol{M}_2$ is assumed to be correct a priori and the red dashed line is the post-selection density of t-statistics $t = \hat{\beta}_1/\hat{\sigma}_{\beta_1}$)

The resulting t-values are shown in Figure 2.2. As we can see, after model selection, the density of t_1 no longer follows t-distribution. The distortion is more severe than in the first case, in the sense that the post-selection density of t-values are now a bimodal distribution. The mean has shifted from approximately 1.90 to 3.05 and standard deviation has shifted from approximate 1.11 to 2.21. Now, confidence interval derived from the t-statistics will not provide valid unconditional error control.

One may hope that the distribution of t-statistics conditional on the the true model $\boldsymbol{M}_2$ being selected remains valid, even though the true model would not be known a priori in practice. However, this expectation can not be satisfied either. We also plot the density of t-statistics conditional on $\boldsymbol{M}_2$ being selected in the Figure 2.3. We can see that the density of t-statistics conditional on selected $\boldsymbol{M}_2$ is still different from the density assuming $\boldsymbol{M}_2$ correct a priori. The post-selection mean and standard deviation are 1.42 and 0.97. Hence, it seems that selecting the true model only ensures that the proper predictors are included. It does not guarantee any of the desirable properties of the coefficient estimates.

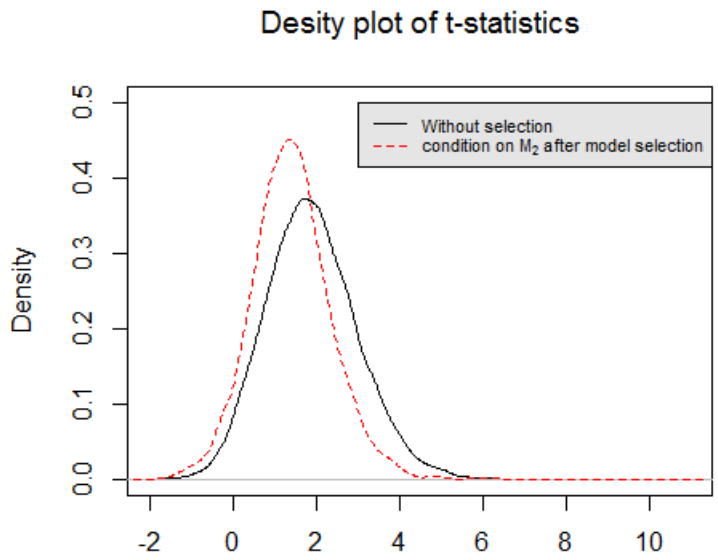

Figure 2.3: The density plot of t-statistics (The solid black line is the density where $\boldsymbol{M}_2$ is assumed to be correct a priori and the dashed red line is the post-selection density of t-statistics $t = \hat{\beta}_1/\hat{\sigma}_{\beta_1}$ conditional on $\boldsymbol{M}_2$ being selected.)

Case 3: Now, we consider the case that $\alpha_1 = 0.5$, $\alpha_2 = 0$ and the variance σ_ε^2 must be estimated. In this case, both $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ are correct in the sense that $\beta_2 = 0$ in $\boldsymbol{M}_2$. Usually, we call $\boldsymbol{M}_1$ as the preferred model or the most parsimonious model, as it only employs one predictor and therefore affords more degrees of freedom for testing.

We plot the density of t-statistics in Figure 2.4. The black solid line represents again the density when $\boldsymbol{M}_1$ is known to be correct a priori and red dashed line represents post-selection sampling distribution. In this case, we have 1866 times $\boldsymbol{M}_2$ selected out of 10000 experiments. The post-selection sampling distribution of t-statistics is still slightly different than the sampling distribution in the case that $\boldsymbol{M}_1$ is assumed to be correct a priori. Although the

post-selection mean is only slightly shifted to 1.50 from 1.60, the 2.5% quantile differs a lot. The post-selection 2.5% quantile is -0.44, and the 2.5% quantile assuming $\boldsymbol{M}_1$ correct is -1.01. This suggests that the post-selection confidence interval constructed based on t-statistics would be invalid.

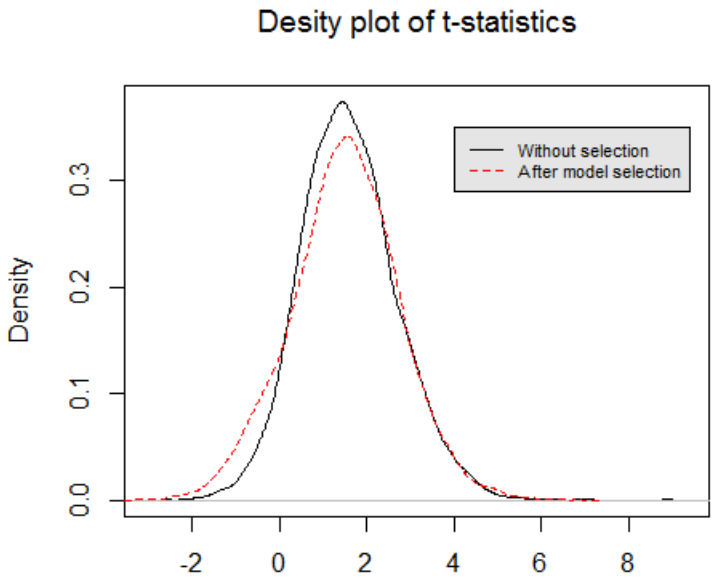

Figure 2.4: The density plot of t-statistics (The solid black line is the density where $\boldsymbol{M}_1$ is assumed to be correct as a priori and the red dashed line is the post-selection density of t-statistics $t = \hat{\beta}_1/\hat{\sigma}_{\beta_1}$).

Chapter 3

Theoretical Assumptions and Target of Inference

3.1 Interpretation of parameters in submodel

Model selection not only raises the problem of post-selection inference, but also raises a problem which is no less important: the meaning and role of parameters in the submodels. There are two different viewpoints regarding this problem, namely full model view and submodel view. In this section, we generally describe these two different views and its corresponding statistical meaning in the linear model.

3.1.1 The full model interpretation

In the full model view of linear regression, coefficients of predictors are always interpreted as full model parameters. The selection of model parameters is interpreted as setting non-selected parameters to zero. Hence, the parameters of submodel are actually parameters of the full model under a zero constraint on the non-selected parameters. Thus, regardless of whether a predictor is selected or not selected, its coefficient estimate always exists.

Underlying the full model view, the full model is regarded as a "data generating" machine. The estimation of coefficients is intended to estimate the underlying full equation which describe the "data generating" mechanism for the response. And thus predictors have a causal interpretation for the response.

3.1.2 The submodel interpretation

In another view, called the submodel view, each submodel has its own parameter space. Non-selected parameters are not zero but do not exist in the corresponding submodel parameter space. An important point which will influence the target of subsequent post-selection inference is that the estimation under submodel view aims to estimate any equation which merely describes associations between the predictor and response variables. Thus, the goal is not to learn about the full model and causal effects.

In Berk et al. (2013), they summarize three points that make the submodel interpretation of coefficients different from the full model view:

1. The full model has no special status other than being the repository of available predictors.

2. The coefficients of excluded predictors are not zero; they are are not defined and therefore do not exist.

3. The meaning of a predictor's coefficient depends on which other predictors are included in the selected model.

The three points importantly influence the validation of the PoSI-intervals, which will be illustrated in the next section.

3.2 Preliminary work

We now layout assumptions and notations in post-selection inference throughout the thesis. The framework generally adopts all necessary background of Berk et al. (2013). In this thesis, we consider the Gaussian response vector $\boldsymbol{Y} \in \mathbb{R}^n$ with $\mathbf{E}[\boldsymbol{Y}] = \boldsymbol{\mu} \in \mathbb{R}^n$ and common variance $\sigma^2 > 0$, that is

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I}_n)$. The normal assumption is needed in the case of testing and constructing CI. Estimation does not require the normality assumption.

The design matrix of full model is denoted as a $n \times p$ matrix $\boldsymbol{X}$, where $p > n$ is actually allowed. To denote a submodel matrix and corresponding predictors, we also follow the convention in the literature. We first denote the full model using index set $\boldsymbol{M}_F = \{1, \ldots, p\}$, and thus $\boldsymbol{X} = \boldsymbol{X}_{\boldsymbol{M}_F}$. To denote a submodel, we use an indexing set $\boldsymbol{M} = \{j_1, j_2, \ldots, j_m\} \subset \boldsymbol{M}_F$ with $\boldsymbol{M} \neq \emptyset$. The size of a submodel is $|\boldsymbol{M}| = m$ and the size of full model is $|\boldsymbol{M}_F| = p$. Thus, we have $\boldsymbol{X}_{\boldsymbol{M}} = (\boldsymbol{X}_{j_1}, \ldots, \boldsymbol{X}_{j_m})$, where the $\boldsymbol{X}_j$ is the jth column of $\boldsymbol{X}$. Write $\mathcal{M}$ as a user-specified (nonempty) collection of candidate models. Throughout, we assume that $\mathcal{M}$ consists only of submodels of full column rank, that is, we assume that the rank of $\boldsymbol{X}_{\boldsymbol{M}}$ equals $|\boldsymbol{M}|$ and satisfies $1 \leq |\boldsymbol{M}| \leq n$ for $\boldsymbol{M} \in \mathcal{M}$.

3.3 Variance estimation

Further, we need to assume that we have an estimator of $\hat{\sigma}^2$ for σ^2. The reason that we need to point out this problem separately is that the first-order correctness is generally not assumed in the framework of PoSI. However, for inference, we need a way to provide a valid estimation for σ^2. The $\hat{\sigma}^2$ should be independent of all least squares estimators in the PoSI procedure introduced shortly. For the estimator $\hat{\sigma}^2$, if the variance is unknown, we will assume it is distributed as a chi-squared random variable with r degrees of freedom multiplied by σ^2/r, that is $\hat{\sigma}^2 \sim \sigma^2\chi^2_r/r$. If the variance σ^2 is known, then we set $\hat{\sigma}^2 = \sigma^2$ and $r = \infty$. The joint distribution of $\boldsymbol{Y}$ and $\hat{\sigma}^2$ depend on the parameter $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\sigma^2 > 0$.

There are several ways to provide this independent variance estimator $\hat{\sigma}^2$:

- In the classical case, the most common way is to assume that the full model is first-order correct, i.e. $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$ in addition to the assumption $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$, which leads to the mean squared residuals $\hat{\sigma}^2_F = \|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|/(n-p)$ as the variance estimator we desired.

- Exact replications of the response may sometimes be obtained under the same conditions. And the estimate $\hat{\sigma}^2$ can be obtained from mean square residuals of the one-way ANOVA of the groups of replicates.

- A random split-sample approach where one part of the data is used for estimating $\hat{\sigma}^2$ and the other part for estimating coefficients, selecting models and making post-selection inference, would generate a variance estimator as desired.

3.4 Submodel coefficient estimation

Under a submodel $\boldsymbol{M} \in \mathcal{M}$, $\boldsymbol{Y}$ is modeled as

$$\boldsymbol{Y} = \boldsymbol{X_M}\boldsymbol{\beta_M} + \boldsymbol{\delta_M}$$

where $\boldsymbol{\beta_M}$ is called regression coefficients of the predictor matrix $\boldsymbol{X_M}$. It corresponds to the orthogonal projection of $\boldsymbol{\mu}$ from onto the column-space of X_M, and is defined as

$$\boldsymbol{\beta_M} \triangleq (\boldsymbol{X_M}^T\boldsymbol{X_M})^{-1}X_M^T\mathbf{E}[\boldsymbol{Y}] = (\boldsymbol{X_M}^T\boldsymbol{X_M})^{-1}\boldsymbol{X_M}^T\boldsymbol{\mu}\,. \tag{3.1}$$

The unique least squares estimator for $\boldsymbol{\beta_M}$ in $\boldsymbol{M}$ is

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{M}} = (\boldsymbol{X_M}^T\boldsymbol{X_M})^{-1}\boldsymbol{X_M}^T\boldsymbol{Y}. \tag{3.2}$$

Notation: To locate the regression coefficients of the predictor X_j relative to the submodel it appears in, we write $\boldsymbol{\beta}_{j\cdot\boldsymbol{M}} = \mathbf{E}[\hat{\boldsymbol{\beta}}_{j\cdot\boldsymbol{M}}]$ for the component of $\boldsymbol{\beta_M} = \mathbf{E}[\hat{\boldsymbol{\beta}}_{\boldsymbol{M}}]$ that corresponds to the regressor $\boldsymbol{X}_j$ for each $j \in \boldsymbol{M}_F = \{1, \ldots, p\}$. This is called "full model indexing" in Berk et al. (2013).

Remark: There are several remarks regarding to the estimation target $\boldsymbol{\beta_M}$ defined above:

1. In the classical case $p \leq n$, we can define the target of the full-model estimator $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{Y}$ as a special case of (3.2) with $\boldsymbol{M} = \boldsymbol{M}_F$, and $\boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathbf{E}[\boldsymbol{Y}]$.

2. In general, let $\boldsymbol{\beta}$ be any minimizer of $\left\|\boldsymbol{\mu} - \boldsymbol{X}\boldsymbol{\beta}^T\right\|^2$. There is a link between $\boldsymbol{\beta}_{\boldsymbol{M}}$ and $\boldsymbol{\beta}$,

$$\boldsymbol{\beta}_{\boldsymbol{M}} = (\boldsymbol{X}_{\boldsymbol{M}}^T \boldsymbol{X}_{\boldsymbol{M}})^{-1} \boldsymbol{X}_{\boldsymbol{M}}^T \boldsymbol{X}\boldsymbol{\beta}.$$

   Thus the target $\boldsymbol{\beta}_{\boldsymbol{M}}$ is an estimable linear function of $\boldsymbol{\beta}$, even without first-order correctness assumptions.

3. If the model $\boldsymbol{M}$ is first-order correct, i.e $\boldsymbol{X}_{\boldsymbol{M}}\boldsymbol{\beta}_{\boldsymbol{M}} = \boldsymbol{\mu}$, then we have $\delta_{\boldsymbol{M}} = \boldsymbol{\varepsilon}$. If the model is not first-order correct, than we have $\boldsymbol{\delta}_{\boldsymbol{M}} = \boldsymbol{\mu} - \boldsymbol{X}_{\boldsymbol{M}}\boldsymbol{\beta}_{\boldsymbol{M}} + \boldsymbol{\varepsilon}$. Regardless of the correctness of the model, under the above normality assumption, we always have $\hat{\boldsymbol{\beta}}_{\boldsymbol{M}} \sim N(\boldsymbol{\beta}_{\boldsymbol{M}}, \sigma^2(\boldsymbol{X}_{\boldsymbol{M}}^T \boldsymbol{X}_{\boldsymbol{M}})^{-1})$. Therefore, Regardless of whether the model $\boldsymbol{M}$ is correct or not, $\hat{\boldsymbol{\beta}}_{\boldsymbol{M}}$ is an unbiased estimator for $\boldsymbol{\beta}_{\boldsymbol{M}}$.

4. The estimation target, $\boldsymbol{\beta}_{\boldsymbol{M}}$ is not the conventional estimation target in the full model view setting, even under first-order correctness assumption. Traditionally, in the full model view, we assume the underlying "data generating" model as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\alpha}^T + \boldsymbol{\varepsilon}$, where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)$. Then submodel inference target in full model viewpoint is $\boldsymbol{\alpha}_{\boldsymbol{M}} = (\alpha_{j_1}, \ldots, \alpha_{j_m})$, where α_j is the jth element of vector $\boldsymbol{\alpha}$. Apparently, $\boldsymbol{\alpha}_M \neq \boldsymbol{\beta}_{\boldsymbol{M}}$ unless the $\boldsymbol{M}$ is the so-called the most parsimonious "true" model, which means that the coefficients not contained in the submodel M are zero, i.e. $\boldsymbol{\alpha}_{\boldsymbol{M}_F \setminus \boldsymbol{M}} = 0$, or $\boldsymbol{X}$ is an orthogonal matrix.

   This point is actually very important, since the validation of PoSI-intervals is established based on this unconventional quantity of interest. Hence, by design, the PoSI-intervals do not provide a solution to the more traditional problem, where the goal is to cover a parameter in the overall model after model selection, i.e. $\alpha_{\boldsymbol{M}}$.

Chapter 4

Valid Post-selection Inference

In this thesis, we mainly consider the post-selection inference problem as building up confidence intervals after model selection. Following the last section, we recall the assumption needed is $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$. Under the assumption, we have that

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{M}} \sim N(\boldsymbol{\beta}_{\boldsymbol{M}}, \sigma^2 (\boldsymbol{X}_{\boldsymbol{M}}^T \boldsymbol{X}_{\boldsymbol{M}})^{-1}), \tag{4.1}$$

or equivalently,

$$\hat{\boldsymbol{\beta}}_{j \cdot \boldsymbol{M}} \sim N(\boldsymbol{\beta}_{j \cdot \boldsymbol{M}}, \sigma^2 / \left\| \boldsymbol{X}_{j \cdot M} \right\|^2). \tag{4.2}$$

In the unknown variance case, again, according to the last section, we assume there is a valid estimate $\hat{\sigma}^2$ of σ^2 that is independent of all estimates $\hat{\beta}_{j \cdot \boldsymbol{M}}$, and we further assume $\hat{\sigma}^2 \sim \sigma \chi_r^2 / r$ with r degrees of freedom. If the full model is assumed to be correct and $n > p$, then $r = n - p$. In the known variance case, $r = \infty$.

In the unknown variance case, without model selection and first-order correctness of submodel $\boldsymbol{M}$, we construct t-statistics $t_{j \cdot \boldsymbol{M}}$ that uses $\hat{\sigma}^2$ instead of $\hat{\sigma}_{\boldsymbol{M}}^2$ for inference on $\beta_{j \cdot \boldsymbol{M}}$ as

$$t_{j \cdot \boldsymbol{M}} \triangleq \frac{\hat{\boldsymbol{\beta}}_{j \cdot \boldsymbol{M}} - \boldsymbol{\beta}_{j \cdot \boldsymbol{M}}}{[(\boldsymbol{X}_{\boldsymbol{M}}^T \boldsymbol{X}_{\boldsymbol{M}})^{-1}]_{jj}^{1/2} \hat{\sigma}} = \frac{\hat{\boldsymbol{\beta}}_{j \cdot \boldsymbol{M}} - \boldsymbol{\beta}_{j \cdot \boldsymbol{M}}}{\hat{\sigma} / \left\| \boldsymbol{X}_{j \cdot \boldsymbol{M}} \right\|} = \frac{(\boldsymbol{Y} - \boldsymbol{\mu})^T \boldsymbol{X}_{j \cdot \boldsymbol{M}}}{\hat{\sigma} \left\| \boldsymbol{X}_{j \cdot \boldsymbol{M}} \right\|} \tag{4.3}$$

It's easy to see that $t_{j\cdot\boldsymbol{M}}$ has a central t-distribution with r degrees of freedom. Now, the confidence intervals for $\beta_{j\cdot\boldsymbol{M}}$ take the form

$$\mathbf{CI}_{j\cdot\boldsymbol{M}}(K) \triangleq \hat{\boldsymbol{\beta}}_{j\cdot\boldsymbol{M}} \pm K[(\boldsymbol{X}_{\boldsymbol{M}}^T\boldsymbol{X}_{\boldsymbol{M}})^{-1}]_{jj}^{1/2}\hat{\sigma} = \hat{\boldsymbol{\beta}}_{j\cdot\boldsymbol{M}} \pm K\hat{\sigma}/\left\|\boldsymbol{X}_{j\cdot\boldsymbol{M}}\right\|. \tag{4.4}$$

K is determined by the user-specified $1-\alpha$ coverage level. In this case, if we choose $K = t_{r,1-\alpha/2}$ to be the $1-\alpha/2$ quantile of a t-distribution with r degrees of freedom, then the interval is marginally valid with a $1-\alpha$ coverage guarantee, namely,

$$\mathbb{P}[\boldsymbol{\beta}_{j\cdot\boldsymbol{M}} \in \mathbf{CI}_{j\cdot\boldsymbol{M}}(K)] \geq 1-\alpha. \tag{4.5}$$

In the known variance case, $t_{j\cdot\boldsymbol{M}}$ turns into a z-statistics which follows standard normal distribution and K turns into the $1-\alpha/2$ quantile of a standard normal distribution, $\Phi^{-1}(1-\alpha/2)$.

4.1 Model selection and its implication for parameters

Now, we consider adding model selection into the framework. Suppose a data-driven model selection procedure is to select a model from the collection of candidate models, $\mathcal{M}$. We denote this mapping as

$$\hat{\boldsymbol{M}} : \boldsymbol{Y} \mapsto \hat{\boldsymbol{M}}(Y), \quad \mathbb{R}^n \to \mathcal{M}, \tag{4.6}$$

where $\hat{\boldsymbol{M}}(\boldsymbol{Y})$ denotes the dependency on the data $\boldsymbol{Y}$.

The resulting post-selection coefficient estimator is $\hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{M}}}$. Now, the $\boldsymbol{\beta}_{\hat{\boldsymbol{M}}}$ is the quantity of interest in post-selection inference. It is a random quantity as it depends on the outcome of the model selection procedure. Both the meaning and the dimension of $\boldsymbol{\beta}_{\hat{\boldsymbol{M}}}$ will be influenced by the data $(\boldsymbol{Y}, \boldsymbol{X})$ and the model selection procedures (like AIC, BIC, the Lasso, etc.). We refer to Berk et al. (2013) for further discussion about the motivation of studying $\boldsymbol{\beta}_{\hat{\boldsymbol{M}}}$, and also to Leeb et al. (2015) for the debate for setting $\boldsymbol{\beta}_{\hat{\boldsymbol{M}}}$ as the inference target.

4.2 Universal post-selection coverage guarantees of confidence intervals

After taking model selection into consideration, the post-selection inference target become $\boldsymbol{\beta}_{\hat{\boldsymbol{M}}}$. Since the inference target becomes random, (4.5) would fail as the stochastic nature of the selection procedure will influence the distribution of t-statistics. In Berk et al. (2013), they propose the PoSI intervals which construct simultaneous confidence intervals for all possible $\boldsymbol{\beta}_{\hat{\boldsymbol{M}}}$, i.e. all $\boldsymbol{\beta}_{\boldsymbol{M}}$ for $\boldsymbol{M} \in \mathcal{M}$, with the universal coverage guarantee $1-\alpha$. Formally speaking, the PoSI-intervals can be constructed by choosing K_p such that

$$\begin{aligned} &\mathbb{P}[\boldsymbol{\beta}_{j\cdot\boldsymbol{M}} \in \mathbf{CI}_{j\cdot\boldsymbol{M}}(K_p) : j \in \boldsymbol{M}, \boldsymbol{M} \in \mathcal{M}] \\ =&\mathbb{P}[\boldsymbol{\beta}_{j\cdot\boldsymbol{M}} \in \hat{\boldsymbol{\beta}}_{j\cdot\boldsymbol{M}} \pm K_p[(\boldsymbol{X}_{\boldsymbol{M}}^T\boldsymbol{X}_{\boldsymbol{M}})^{-1}]_{jj}^{1/2}\hat{\sigma} : j \in \boldsymbol{M}, \boldsymbol{M} \in \mathcal{M}] \\ \geq&1-\alpha. \end{aligned} \tag{4.7}$$

The K_p is called as PoSI-constant for convenience. Note that (4.7) is equivalent to

$$\mathbb{P}[\boldsymbol{\beta}_{j\cdot\hat{\boldsymbol{M}}} \in \mathbf{CI}_{j\cdot\hat{\boldsymbol{M}}}(K_p) : \forall j \in \hat{\boldsymbol{M}}] \geq 1-\alpha \qquad \forall \hat{\boldsymbol{M}} \in \mathcal{M}. \tag{4.8}$$

There are several remarks about PoSI-intervals:

1. The coverage guarantee of PoSI-intervals is universally valid in post-selection inference, namely for all model selection procedures $\hat{\boldsymbol{M}}$. The "universal" guarantee also implies a "family-wise" coverage guarantee for all selected predictor $j \in \hat{\boldsymbol{M}}$, i.e.

$$\mathbb{P}[\boldsymbol{\beta}_{j\cdot\hat{\boldsymbol{M}}} \in \mathbf{CI}_{j\cdot\hat{\boldsymbol{M}}}(K_p) : \forall j \in \hat{\boldsymbol{M}}] \geq 1-\alpha, \tag{4.9}$$

   although the "family-wise" is unusual as $\hat{\boldsymbol{M}}$ is random.

2. As the PoSI guarantee is both universal and "familywise", we can say something very strong, regardless of realization $\boldsymbol{y}$ of $\boldsymbol{Y}$: "we have $1-\alpha$ confidence that, for $j \in \hat{\boldsymbol{M}}$, the interval $\mathbf{CI}_{j\cdot\hat{\boldsymbol{M}}}(K_p)$ contains $\boldsymbol{\beta}_{j\cdot\hat{\boldsymbol{M}}}$ for any $\boldsymbol{y}$".

3. The universal validity raises question on whether the PoSI-constant is too conservative. We refer to Berk et al. (2013) to further discussion about this.

4. The probability in (4.7) is not hard to compute as it only relies on the random variables $\frac{\hat{\boldsymbol{\beta}}_{j\cdot\boldsymbol{M}}-\boldsymbol{\beta}_{j\cdot\boldsymbol{M}}}{[(\boldsymbol{X}_{\boldsymbol{M}}^T\boldsymbol{X}_{\boldsymbol{M}})^{-1}]_{jj}^{1/2}\hat{\sigma}}$, which follow dependent t-distributions for $j \in \boldsymbol{M}$ and $\boldsymbol{M} \in \mathcal{M}$ in the unknown-variance case and follow dependent normal distributions in the known-variance case. The dependency only relies on $\boldsymbol{X}$.

5. The PoSI-constant in Berk et al. (2013) that makes the guarantee (4.7) hold is designed to make (4.9) hold universally and is immune to the selection procedure. But actually, if the selection procedure is specified, there may be a smaller K' satisfying (4.9). Hence, different selection procedures would require different constants K'. Although it is a possible improvement, it is typically hard to find the procedure-specific constants.

6. In particular, (4.5) holds if K is replaced by K_p for a given $\boldsymbol{M}$.

4.3 One primary predictor and controls-"PoSI1"

Sometimes in the regression, we might have a specific predictor of interest $\boldsymbol{X}_j$ and we want to make inference on $\beta_{j\cdot\hat{\boldsymbol{M}}}$. In this section, suppose j is the index of a fixed and a priori chosen predictor. The other predictors in X are open to be selected. Hence, the user-specified candidate model pool $\mathcal{M}$ becomes a sub-universe $\mathcal{M}_{j\cdot}$ of submodels where $\boldsymbol{X}_j$ is forced to be in all submodel, i.e.

$$\mathcal{M}_{j\cdot} \triangleq \{\boldsymbol{M} | j \in \boldsymbol{M} \in \mathcal{M}\}. \tag{4.10}$$

In such a situation, we only consider constructing valid confidence intervals after model selection for $\beta_{j\cdot\hat{\boldsymbol{M}}}$. Berk et al. (2013) proposes a modification $K_{j\cdot}$ of K_p to construct PoSI1-intervals, such that

$$\mathbb{P}[\boldsymbol{\beta}_{j\cdot\boldsymbol{M}} \in \mathbf{CI}_{j\cdot\boldsymbol{M}}(K_{j\cdot}) : \boldsymbol{M} \in \mathcal{M}_{j\cdot}] \geq 1 - \alpha. \tag{4.11}$$

There are also some remarks regarding this situation:

1. The PoSI1 constants are smaller than PoSI constant, namely, $K_{j\cdot} \leq K_p$ for all j.

2. Generally, when making inference on one primary parameter $\beta_{j\cdot}$, there is a big problem of incoherency. If all predictors in $\boldsymbol{X}$ are open to selected, it makes $\beta_{j\cdot\hat{\boldsymbol{M}}} \in \mathbf{CI}_{j\cdot\hat{\boldsymbol{M}}}(K)$ an incoherent statement that does not even define an event because $\beta_{j\cdot\hat{\boldsymbol{M}}}$ does not exist for $j \notin \hat{\boldsymbol{M}}$. Therefore, the post-selection inference target is not well-defined. PoSI1-intervals bypass this problem by forcing a parameter fixed in all candidates models. Therefore, the inference target, the corresponding coefficient of primary predictor, is well-defined after model selection.

3. Instead of forcing a predictor to be immune from selection, we may also consider the marginal and conditional probabilities

$$\mathbb{P}[j \in \hat{\boldsymbol{M}} \text{ and } \boldsymbol{\beta}_{j\cdot\hat{\boldsymbol{M}}} \in \mathbf{CI}_{j\cdot\hat{\boldsymbol{M}}}(K)] \tag{4.12}$$

and

$$\mathbb{P}[\boldsymbol{\beta}_{j\cdot\hat{\boldsymbol{M}}} \in \mathbf{CI}_{j\cdot\hat{\boldsymbol{M}}}(K) | j \in \hat{\boldsymbol{M}}]. \tag{4.13}$$

These probabilities are both well-defined and can be targets of coverage guarantees. Berk et al. (2013) also provides a coverage guarantee when using PoSI1-constants in these two probabilities,

$$\mathbb{P}[j \in \hat{\boldsymbol{M}} \,\&\, \boldsymbol{\beta}_{j\cdot\boldsymbol{M}} \in \mathbf{CI}_{j\cdot\hat{\boldsymbol{M}}}(K_{j\cdot})] \geq \mathbb{P}[j \in \hat{\boldsymbol{M}}] - \alpha \tag{4.14}$$

and

$$\mathbb{P}[\boldsymbol{\beta}_{j\cdot\boldsymbol{M}} \in \mathbf{CI}_{j\cdot\hat{\boldsymbol{M}}}(K_{j\cdot}) | j \in \hat{\boldsymbol{M}}] \geq 1 - \frac{\alpha}{\mathbb{P}[j \in \hat{\boldsymbol{M}}]}. \tag{4.15}$$

We refer to Appendix B.4 of Berk et al. (2013) for further proof and discussion.

Chapter 5

Simulation Study

In this section, we want to carry out a simulation study to investigate the performance of different types of confidence intervals, i.e. the minimal coverage probabilities. We follow the approach of Leeb et al. (2015). However, we improve the accuracy, adjusted to the simulation setting.

5.1 Data generating model and target of inference

In our simulation study, the data is generated from a Gaussian linear model which takes the form

$$\boldsymbol{Y}_{n\times 1} = \boldsymbol{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}, \tag{5.1}$$

where $n = 30$, $p = 10$ and $\boldsymbol{\varepsilon}_{n\times 1} \sim N(\mathbf{0}, \boldsymbol{I}_n)$. Hence the full model $\boldsymbol{M}_F = \{1, \ldots, 10\}$. Denote $\boldsymbol{y}$ as a realization of $\boldsymbol{Y}$.

For the variance estimator $\hat{\sigma}^2$, we use the usual unbiased variance estimator of residual sum square by assuming first-order correctness and fitting the full model. Therefore, $\hat{\sigma}^2 \sim \sigma^2\chi_r/r$ with $r = n - p = 20$ degrees of freedom.

The inference target is the coefficient corresponding to $\boldsymbol{X}_1$. The pool of candidate models become

$$\mathcal{M}_{1\cdot} \triangleq \{\boldsymbol{M} | 1 \in \boldsymbol{M} \in \mathcal{M}\}. \tag{5.2}$$

We consider the random inference target $\beta_{1\cdot\hat{\boldsymbol{M}}} = [(\boldsymbol{X}_{\hat{\boldsymbol{M}}}^T \boldsymbol{X}_{\hat{\boldsymbol{M}}})^{-1} \boldsymbol{X}_{\hat{\boldsymbol{M}}}^T \boldsymbol{\mu}]_1$ as well as the conventional inference target $\boldsymbol{\beta}_1$.

5.2 Design matrix

Three design matrices are considered in the simulation: For design 1, we take the regressor matrix from the example from page 179 of Rawlings, Pantula, and Dickey (1998). The data is simulated data on peak rate of flow Q (cfs) of water from six watersheds following storm episodes. The storm episodes have been chosen from a larger data set to give a range of storm intensities. The independent variables are

X_1 = Area of watershed (mi^2)

X_2 = Area impervious to water (mi^2)

X_3 = Average slope of watershed (percent)

X_4 = Longest stream flow in watershed (thousands of feet)

X_5 = Surface absorbency index, 0 = complete absorbency, 100 = no absorbency

X_6 = Estimated soil storage capacity (inches of water)

X_7 = Infiltration rate of water into soil (inches/hour)

X_8 = Rainfall (inches)

X_9 = Time period during which rainfall exceeded $\frac{1}{4}$ inch/hr.

The data is shown in Table 5.1.

Table 5.1: Peak flow data from six watersheds.

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	Q
0.03	0.006	3.0	1	70	1.5	0.25	1.75	2	46
0.03	0.006	3.0	1	70	1.5	0.25	2.25	3.7	28
0.03	0.006	3.0	1	70	1.5	0.25	4.00	4.2	54
0.03	0.021	3.0	1	80	1.0	0.25	1.6	1.5	70
0.03	0.021	3.0	1	80	1.0	0.25	3.1	4.0	47
0.03	0.021	3.0	1	80	1.0	0.25	3.6	2.4	112
0.13	0.005	6.5	2	65	2.0	0.35	1.25	0.7	398
0.13	0.005	6.5	2	65	2.0	0.35	2.3	3.5	98
0.13	0.005	6.5	2	65	2.0	0.35	4.25	4.0	191
0.13	0.008	6.5	2	68	0.5	0.15	1.45	2.0	171
0.13	0.008	6.5	2	68	0.5	0.15	2.6	4.0	150
0.13	0.008	6.5	2	68	0.5	0.15	3.9	3.0	331
1.00	0.023	15.0	10	60	1.0	0.2	0.75	1.0	772
1.00	0.023	15.0	10	60	1.0	0.2	1.75	1.5	1,268
1.00	0.023	15.0	10	60	1.0	0.2	3.25	4.0	849
1.00	0.023	15.0	10	65	2.0	0.2	1.8	1.0	2,294
1.00	0.023	15.0	10	65	2.0	0.2	3.1	2.0	1,984
1.00	0.023	15.0	10	65	2.0	0.2	4.75	6.0	900
3.00	0.039	7.0	15	67	0.5	0.5	1.75	2.0	2,181
3.00	0.039	7.0	15	67	0.5	0.5	3.25	4.0	2,484
3.00	0.039	7.0	15	67	0.5	0.5	5.0	6.5	2,450
5.00	0.109	6.0	15	62	1.5	0.6	1.5	1.5	1,794
5.00	0.109	6.0	15	62	1.5	0.6	2.75	3.0	2,067
5.00	0.109	6.0	15	62	1.5	0.6	4.2	5.0	2,586
7.00	0.055	6.5	19	56	2.0	0.5	1.8	2.0	2,410
7.00	0.055	6.5	19	56	2.0	0.5	3.25	4.0	1,808
7.00	0.055	6.5	19	56	2.0	0.5	5.25	6.0	3,024
7.00	0.063	6.5	19	56	1.0	0.5	1.25	2.0	710
7.00	0.063	6.5	19	56	1.0	0.5	2.9	3.4	3,181
7.00	0.063	6.5	19	56	1.0	0.5	4.76	5.0	4,279

For design 2, we consider the exchangeable design in Section 6.1. In exchangeable designs all pairs of predictor vector enclose the same angle. In canonical coordinates, a convenient way to parametrize an exchangeable design is

$$\boldsymbol{X}^{(p)}(a) = \boldsymbol{I}_p + a\boldsymbol{E}_{p\times p} \tag{5.3}$$

where $-1/p < a < \infty$, and $\boldsymbol{E}_{p\times p}$ is a matrix with all entries equal to 1. The range on a is designed to assure $\boldsymbol{X}^{(p)}$ to be positive definite. We choose $a = 10$ in the simulation study, and set $\boldsymbol{X} = U\boldsymbol{X}^{(p)}(a)$. The matrix U is a $n \times p$ orthogonal matrix.

For design 3, we consider the equicorrelated design studied in Section 6.2 of Berk et al. (2013). Similarly, we also have a $p \times p$ design matrix in canonical coordinate:

$$\boldsymbol{X}^{(p)}(c) = (\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_{p-1}, \boldsymbol{X}_p(c)), \tag{5.4}$$

where $\boldsymbol{X}_p(C) = (c, c, \ldots, c, \sqrt{1-(p-1)c^2})^T \in \mathbb{R}^T$ and $\boldsymbol{e}_i \in \mathbb{R}^p$ is canonical coordinates with i-th element to be 1. Then we set $c = \sqrt{0.8/(p-1)}$, and we set $\boldsymbol{X} = V\boldsymbol{X}^{(p)}(c)$, where V is the same type of matrix as U, i.e. $n \times p$ and orthogonal.

5.3 Model selection procedures

As for model selection procedures, we consider AIC, BIC and Lasso. All are implemented in `R`. For AIC, we use the `step` function in `R` with a constraint the $\boldsymbol{X}_1$ is always included. Then the `step` function should search over the 2^9 candidate models of $\mathcal{M}_{1\cdot}$ to minimize AIC penalty function. Similarly, for BIC we also use the `step()` function with the penalty

equal to log(30). For Lasso, we treat lasso as a model selection procedure by selecting variables if the lasso estimator has nonzero coefficient. The Lasso is implemented by `lars` package. To protect the $\boldsymbol{X}_1$ against model selection, we first compute the residual vector $\tilde{\boldsymbol{y}}$ of the orthogonal projection of $\boldsymbol{y}$ on $\boldsymbol{X}_1$.Then we compute Lasso-estimator for the regression between $\tilde{\boldsymbol{y}}$ and $\tilde{\boldsymbol{X}} = (\boldsymbol{X}_2, \ldots, \boldsymbol{X}_p)$. The Lasso-penalty parameter λ is chosen by 10-fold cross validation using the `cv.lars()` function. In both two functions, we need to set the `intercept` parameter to `FALSE` to avoid adding intercept automatically. After obtaining the selected regressor in $\tilde{\boldsymbol{X}}$, we add the first column and refit the selected model with $\boldsymbol{y}$.

5.4 Confidence Intervals

In this simulation, we recall that the target of inference is the coefficient corresponding to predictor X_1, the conventional $\boldsymbol{\beta}_1$ and unconventional $\boldsymbol{\beta}_{1\cdot\hat{\boldsymbol{M}}}$. The user-specified $\mathcal{M}_{1\cdot}$ satisfies that $1 \in \boldsymbol{M}, \forall M \in \mathcal{M}_{1\cdot}$. Taking $\beta_{1\cdot\hat{\boldsymbol{M}}}$ for example, we wish to construct confidence intervals for $\beta_{1\cdot\hat{\boldsymbol{M}}}$ which have the form of

$$\hat{\boldsymbol{\beta}}_{1\cdot\hat{\boldsymbol{M}}} \pm K\hat{\sigma}_{1\cdot\hat{\boldsymbol{M}}} \tag{5.5}$$

for some constant $K > 0$, with $\hat{\sigma}^2_{1\cdot\boldsymbol{M}}$ defined by $\hat{\sigma}^2_{1\cdot\boldsymbol{M}} = \hat{\sigma}^2[(\boldsymbol{X}_{\boldsymbol{M}}^T\boldsymbol{X}_{\boldsymbol{M}})^{-1}]_{1,1}$. For a given level of $1-\alpha$ with $0 < \alpha < 1$, the constant K should be chosen such that the minimal coverage probability is at least $1-\alpha$, namely,

$$\mathbb{P}[\boldsymbol{\beta}_{1\cdot\hat{\boldsymbol{M}}} \in \hat{\boldsymbol{\beta}}_{1\cdot\hat{\boldsymbol{M}}} \pm K\hat{\sigma}_{1\cdot\hat{\boldsymbol{M}}}] \geq 1-\alpha. \tag{5.6}$$

The first type of confidence interval is called "naive" interval in the sense the confidence interval construction ignores the model selection. Therefore, then we treat $(\hat{\boldsymbol{\beta}}_{1\cdot\hat{\boldsymbol{M}}} - \boldsymbol{\beta}_{1\cdot\hat{\boldsymbol{M}}})/\hat{\sigma}_{1\cdot\hat{\boldsymbol{M}}}$ as

standard normal distribution in the known-variance case and t-distribution with r degrees of freedom in the unknown-variance case. Then we can find the K_N as $(1-\alpha/2)$-quantile of standard normal distribution in the known-variance case and $(1-\alpha/2)$-quantile of t-distribution with r degrees of freedom in the known-variance case. Then the "naive" interval is

$$\hat{\boldsymbol{\beta}}_{1\cdot\hat{\boldsymbol{M}}} \pm K_N \hat{\sigma}_{1\cdot\hat{\boldsymbol{M}}}. \tag{5.7}$$

The second type and third type of confidence intervals are PoSI-intervals and PoSI1-intervals as we described in Chapter 4. The PoSI-constant K_P now satisfies

$$\mathbb{P}[\boldsymbol{\beta}_{j\cdot\boldsymbol{M}} \in \hat{\boldsymbol{\beta}}_{j\cdot\boldsymbol{M}} \pm K_P \hat{\sigma}_{j\cdot\boldsymbol{M}} : j \in \boldsymbol{M}, \boldsymbol{M} \in \mathcal{M}_{1\cdot}] \geq 1-\alpha, \tag{5.8}$$

and PoSI1-constant K_{P1} satisfies

$$\mathbb{P}[\boldsymbol{\beta}_{1\cdot\boldsymbol{M}} \in \hat{\boldsymbol{\beta}}_{1\cdot\boldsymbol{M}} \pm K_P \hat{\sigma}_{1\cdot\boldsymbol{M}} : \boldsymbol{M} \in \mathcal{M}_{1\cdot}] \geq 1-\alpha. \tag{5.9}$$

So the PoSI-interval is

$$\hat{\boldsymbol{\beta}}_{1\cdot\hat{\boldsymbol{M}}} \pm K_P \hat{\sigma}_{1\cdot\hat{\boldsymbol{M}}}, \tag{5.10}$$

and PoSI1-interval is

$$\hat{\boldsymbol{\beta}}_{1\cdot\hat{\boldsymbol{M}}} \pm K_{P1} \hat{\sigma}_{1\cdot\hat{\boldsymbol{M}}}. \tag{5.11}$$

The last type of intervals are called Scheffé intervals. The Scheffé constant K_S is chosen such that

$$\mathbb{P}\left[\sup_{\nu\in\text{span}(X),\nu\neq 0} \frac{\nu'(Y-\nu)}{\hat{\sigma}\,\|\nu\|} \leq K_S\right] = 1-\alpha. \tag{5.12}$$

Four reference, we refer to the four constants K_N, K_P, K_{P1}, K_S as adjusting constants. The four constants satisfy that $K_N \leq K_{P1} \leq K_P \leq K_S$ by construction. The K_P, K_{P1}

and K_S can provide valid post-selection coverage guarantee because these constants provide simultaneous confidence intervals for all target quantities that can occur after model selection.

5.5 Simulation procedure

The simulation procedure to estimate the minimal coverage probability is comparable with the approach of Leeb et al. (2015). For all three matrices, we simulate realizations $\boldsymbol{y}$ of $\boldsymbol{Y}$ under the model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ for randomly selected values of $\boldsymbol{\beta}$ from standard normal distribution. Among these $\boldsymbol{\beta}$ values, we identify those values for which the simulated coverage probability gets small and also correct the bias.

For example, suppose that we want to investigate in the minimal coverage probability of "naive" confidence interval with coverage target $\boldsymbol{\beta}_1$ and model selector AIC. We first select 10,000 realization values of $\boldsymbol{\beta}$ by drawing i.i.d samples from a random p-dimensional vector $\boldsymbol{b}$ which follows standard normal distribution, namely, $\boldsymbol{b} \sim N(\boldsymbol{0}, \boldsymbol{I})$. After set $\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{b}$, then $\boldsymbol{\beta}$ is our desired samples of $\boldsymbol{\beta}$. Or we can directly draw samples from $\boldsymbol{\beta} \sim N(\boldsymbol{0}, (\boldsymbol{X}^T\boldsymbol{X})^{-1})$. In our experiment, we favor the later approach.

For each of these $\boldsymbol{\beta}$'s, we generate 100 Monte Carlo samples $\boldsymbol{y}^{(i)}_{n\times 1}$ of response $\boldsymbol{Y}$ by drawing the random noise vector $\boldsymbol{\varepsilon}^{(i)}_{n\times 1}$ and setting $\boldsymbol{y}^{(i)}_{n\times 1} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}^{(i)}_{n\times 1}$, $i = 1, \ldots, 100$. Then we approximate the corresponding coverage probability by the coverage rate obtained from the Monte Carlo samples. For each sample $\boldsymbol{y}^{(i)}_{n\times 1}$, we compute the model selector $\hat{\boldsymbol{M}}$ and construct the "naive" confidence interval in the selected model $\hat{\boldsymbol{M}}$. After checking whether β_1 is covered or not, we take the average value of 100 results. Then the average value results

in a coverage rate which is an estimate for the coverage probability of the "naive" interval for the true value β_1.

After repeating this for each of the 10,000 $\boldsymbol{\beta}$'s, we compute the resulting smallest coverage rate as an estimator for the minimal coverage probability of the "naive" confidence interval. This is notably biased downward. (The bias can be explained by the example of the first order statistics of i.i.d. samples from uniform distribution. Although each sample is an unbiased estimator for mean 1/2, the expectation of first order statistics is $1/(1+n)$, much less than its mean). To correct the bias, we select out the $\boldsymbol{\beta}$'s which achieve the smallest 100 coverage rates. Then, for each of these 100 $\boldsymbol{\beta}$'s, we now use 500 Monte Carlo samples to estimate the coverage probability in the similar way above. Then we obtain the $\boldsymbol{\beta}$ which gives the smallest coverage rate. As a third step, we now use 100,000 Monte Carlo samples to get a reliable estimate of the corresponding minimal coverage probability.

In summary, the procedure above is used to compute estimates for the minimal coverage probability of the 72 combinations of different design matrices (three designs as described in Section 5.2), different model selectors (AIC, BIC and Lasso), different coverage targets ($\boldsymbol{\beta}$ and $\boldsymbol{\beta}_{\hat{\boldsymbol{M}}}$) and different types of confidence intervals ("naive", PoSI, PoSI1 and Scheffé).

5.6 Result Analysis

We present the simulation results in this section. We first show that the adjusting constants, i.e. K_N, K_P, K_{P1}, K_S, for different design matrices in Table 5.2. We can see that these constants satisfies that $K_N < K_{P1} < K_P < K_S$ for all three designs. Further, generally speaking, the K_P and K_S are much larger then K_N because they not only consider the

multiplicity (i.e. consider all parameters within $\boldsymbol{M}$) but also consider the universality (i.e. consider all possible $\boldsymbol{M}$).

5.6.1 Minimal coverage probability for $\boldsymbol{\beta}_{1\cdot\hat{\boldsymbol{M}}}$

Table 5.2: Adjusting constants for different types of confidence intervals and different design matrices

Confidence interval	**Design 1 (watershed)**	**Design 2 exchangeable**	**Design 3 equicorr.**
"naive"	2.086	2.086	2.086
PoSI	3.761	3.799	3.817
PoSI1	3.430	3.171	2.710
Scheffé	4.641	4.845	4.845

In Table 5.3, we summarize the smallest coverage probabilities for the unconventional coverage target $\boldsymbol{\beta}_{1\cdot\boldsymbol{M}}$ as proposed in Berk et al. (2013), and, in Table 5.4, we summarize those for the conventional coverage target $\boldsymbol{\beta}_1$.

We first look at Table 5.3 to see minimal coverage probabilities when $\beta_{1\cdot\hat{\boldsymbol{M}}}$ is the inference target. As we expected, when using AIC and BIC as model selector, the "naive" intervals can not provide nominal 0.95 coverage guarantee. By contrast, the PoSI, PoSI1, and Scheffé confidence intervals correct this under-coverage problem. But the correction seems too large as they all result in an over-coverage problem, i.e., the coverage probability are much larger than 0.95. This conforms with the Remark 3 in the Section 4.2 that PoSI and PoSI1 intervals are too conservative.

Table 5.3: Smallest coverage probabilities found in the simulation study for the coverage target $\boldsymbol{\beta}_{1\cdot\hat{\boldsymbol{M}}}$, Using AIC, BIC, Lasso as model selector and four types of confidence intervals including "naive", PoSI, PoSI1 and Scheffé confidence intervals with nominal coverage probability 0.95

Coverage target	**Model selector**	**Confidence interval**	**Design 1 (watershed)**	**Design 2 exchangeable**	**Design 3 equicorr.**
$\boldsymbol{\beta}_{1\cdot\hat{\boldsymbol{M}}}$	AIC	Naive	0.767	0.938	0.940
		PoSI	0.992	0.998	0.998
		PoSI1	0.992	0.998	0.998
		Scheffé	0.999	1.000	1.000
	BIC	Naive	0.779	0.902	0.935
		PoSI	0.989	0.997	0.998
		PoSI1	0.981	0.991	0.984
		Scheffé	0.999	1.000	1.000
	Lasso	Naive	0.959	0.959	0.948
		PoSI	0.999	0.999	0.999
		PoSI1	0.998	0.997	0.987
		Scheffé	1.000	1.000	1.000

What kind of coefficient β_1 results in the smallest coverage probabilities of "naive" intervals? For illustration, we plot the distribution of 100 coefficients β_1 which resulting 100 smallest coverage probabilities of "naive" intervals in the first round of Monte Carlo study in the case of AIC and BIC as model selector for the Design 1 in the Figure 5.6.1. We can see, for design 1, the worst cases are mostly concentrated around -0.5 in both situations. For more similar plots, see Appendix A.

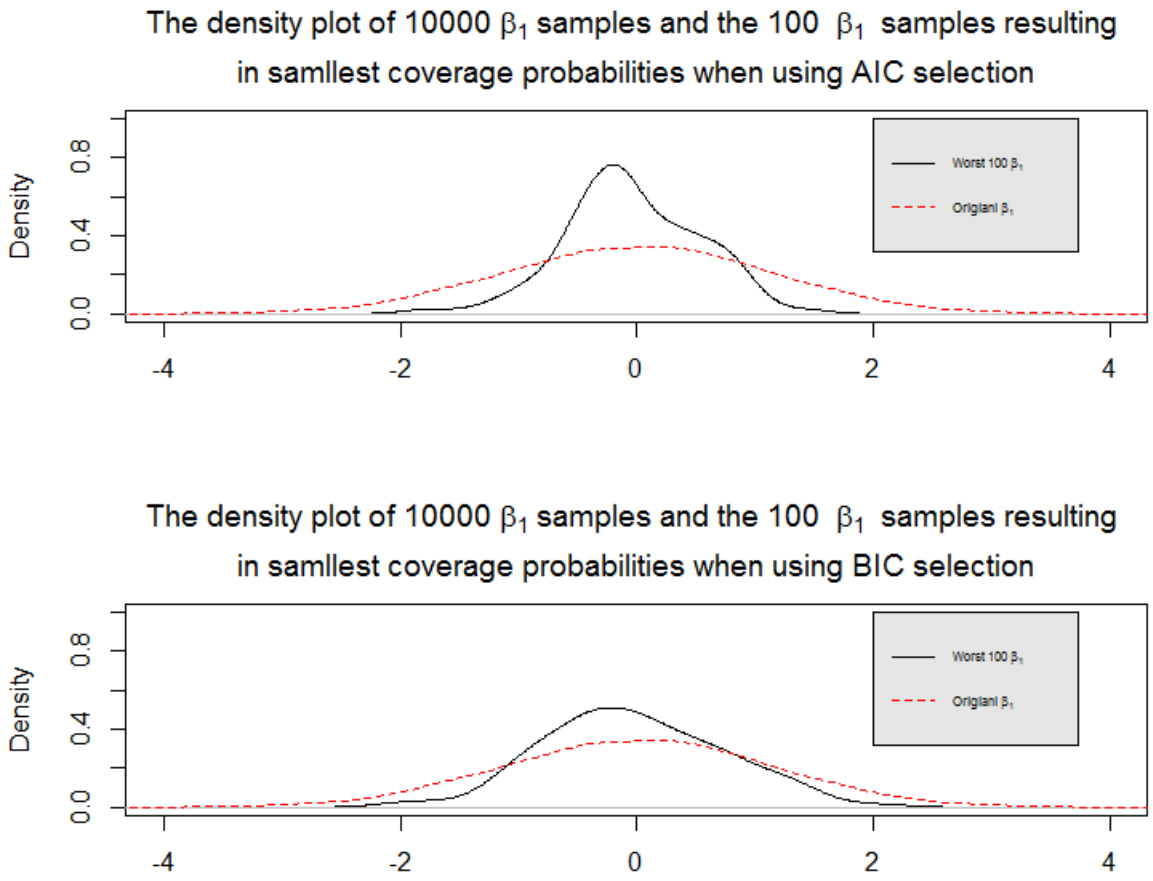

Figure 5.1: The density plot of 100 coefficients β_1 which result in 100 smallest coverage probabilities of "naive" intervals for the Design 1

Notice that, when using Lasso, minimal coverage probability of "naive" intervals is very close to 0.95, which seems to be able to provide the desired nominal 0.95 coverage guarantee. Leeb et al. (2015) suggest an explanation for this phenomenon as quoted below:

"The reason for this is that the LASSO model selector, as implemented here and for the parameters used in the stochastic search for the smallest coverage probability, selects the smallest possible model in most cases, that is, the model containing only the first regressor. In other words, the model selected by the LASSO is nearly nonrandom. Then the target is $\boldsymbol{\beta}_{1\cdot\hat{M}}$, this entails that the naive interval is approximately valid and that both PoSI intervals are too large. [Indeed, the naive interval is valid if the underlying model selector always chooses a fixed (nonrandom) model; cf. the discussion following (2.2).]".

However, under our scrutiny, we actually find that the smallest model, as defined in the above quotation, is only selected in a small portion (less than 20%). This falsifies the explanation in Leeb et al. (2015). Actually, in our opinion, their explanation is counter-intuitive, and we cannot find any reason to explain why the Lasso will select the smallest possible model with high probability, as they suggested. After we carefully check the possible reasons which may lead to this result, we find that the actual explanation comes from the forcing method used to protect $\boldsymbol{X}_1$ from selection as described in Section 5.3.

In the protecting method, the first step is to project the $\boldsymbol{y}$ on the $\boldsymbol{X}_1$ and calculate the residual vector $\tilde{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}_1(\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{y}$. The Lasso selection is based on $\tilde{\boldsymbol{y}}$ and $\tilde{\boldsymbol{X}} = (\boldsymbol{X}_2, \ldots, \boldsymbol{X}_p)$ which deselects the first column $\boldsymbol{X}_1$. Note that the protecting method not only protects the effect of $\mathbf{E}[\boldsymbol{y}] = \boldsymbol{X}^T\boldsymbol{\beta}$ which can be explained by $\boldsymbol{X}_1$, i.e. $\boldsymbol{X}_1^T\boldsymbol{\beta}_1$, but also protects the effect and normality of the random error $\boldsymbol{\varepsilon}$ in the direction of $\boldsymbol{X}_1$. Because of the protection, the model selection does not try to select predictors from $\tilde{\boldsymbol{X}}$ to account for the effect of $\mathbf{E}[\boldsymbol{y}]$ in the direction of $\boldsymbol{X}_1$, and would not be affected by the effect of random error $\boldsymbol{\varepsilon}$ in the direction

of $\boldsymbol{X}_1$ either. Hence, afterwards, when we again regress $\boldsymbol{y}$ on the selected predictors from $\tilde{\boldsymbol{X}}$ together with $\boldsymbol{X}_1$, the "naive" confidence intervals remain valid for $\boldsymbol{X}_1$ because the effects of both $\mathbf{E}[\boldsymbol{y}]$ and $\boldsymbol{\varepsilon}$ in the direction of $\boldsymbol{X}_1$ now remain to be accounted for by $\boldsymbol{X}_1$.

This discovery is quite interesting, because it suggests a possible way to validate "naive" intervals within the PoSI framework in the situation where one primary predictor is the quantity of interest. Although we haven't checked theoretically whether the validation is true and we are not aware of a related theory supporting the validation, we thinks this is quite promising based on the results of simulation study. We also test the minimal coverage probabilities in the case of AIC and BIC as model selectors. The simulated minimal coverage probabilities are all very close to the desired $1 - \alpha$. This validation is very attractive as it may provide exact coverage control, compared to PoSI1-intervals which always result in over-coverage. The theoretical work to check the validation remains as a future work.

5.6.2 Minimal coverage probability for $\boldsymbol{\beta}_1$

We now look at Table 5.3 to see minimal coverage probabilities when we are interested in the conventional inference target $\boldsymbol{\beta}_1$. Most of the confidence intervals fail to provide the coverage guarantee for $\boldsymbol{\beta}_1$. But our experiment result differs from the results of Leeb et al. (2015) where the minimal coverage probabilities of all types of confidence intervals are below 0.95 when the coverage target is $\boldsymbol{\beta}_1$. In our experiment, the Scheffé interval still provide over-coverage for an equicorrelated design matrix (Design 3). But generally speaking, the minimal coverage probabilities of the four types of confidence intervals are sort of random (may depend on the design matrix) and can no longer be above 0.95 in most cases after model selection. Even if using the Lasso selector, we have a very bad minimal coverage rate.

This shows that there is no hope if one wants to make confidence guarantee on the true parameter $\boldsymbol{\beta}$ after model selection.

Table 5.4: Smallest coverage probabilities found in the simulation study for the coverage target $\boldsymbol{\beta}_1$, Using AIC, BIC, Lasso as model selector and four types of confidence intervals including "naive", PoSI, PoSI1 and Scheffé confidence intervals with nominal coverage probability 0.95

Coverage target	**Model selector**	**Confidence interval**	**Design 1 (watershed)**	**Design 2 exchangeable**	**Design 3 equicorr.**
$\boldsymbol{\beta}_1$	AIC	Naive	0.440	0.913	0.893
		PoSI	0.554	0.934	0.996
		PoSI1	0.554	0.934	0.996
		Scheffé	0.593	0.904	1.000
	BIC	Naive	0.214	0.529	0.791
		PoSI	0.316	0.672	0.982
		PoSI1	0.304	0.651	0.913
		Scheffé	0.348	0.650	1.000
	Lasso	Naive	0.108	0.069	0.459
		PoSI	0.132	0.083	0.865
		PoSI1	0.149	0.082	0.630
		Scheffé	0.160	0.085	0.966

Appendix A

The Collection of Plots

In this Appendix, we include two more density plots similar to Figure 5.6.1 in the case of design matrix 2 and design matrix 3.

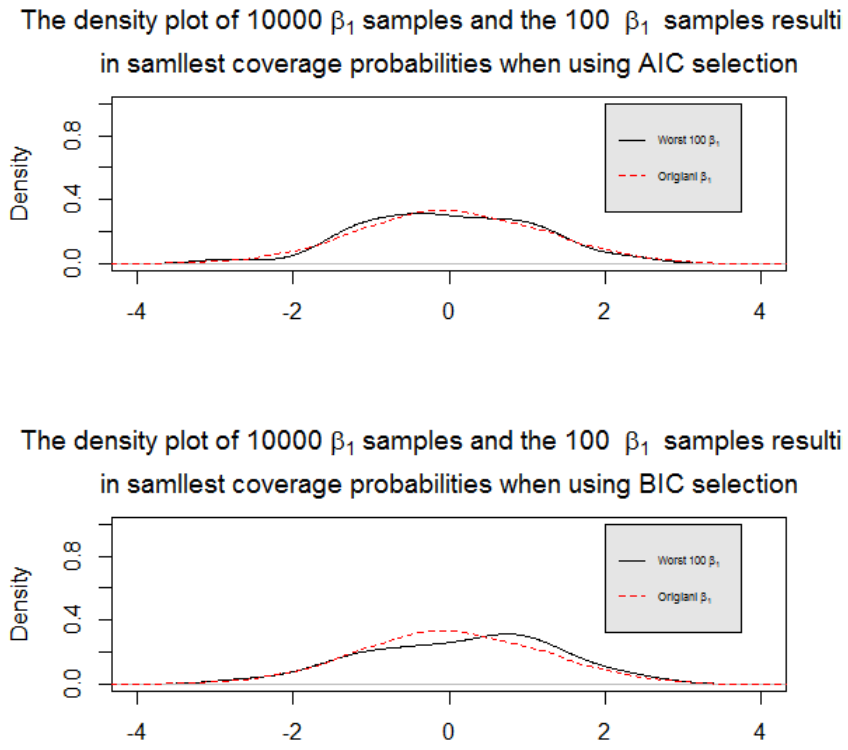

Figure A.1: The density plot of 100 coefficients β_1 which result in 100 smallest coverage probabilities of "naive" intervals for Design 2

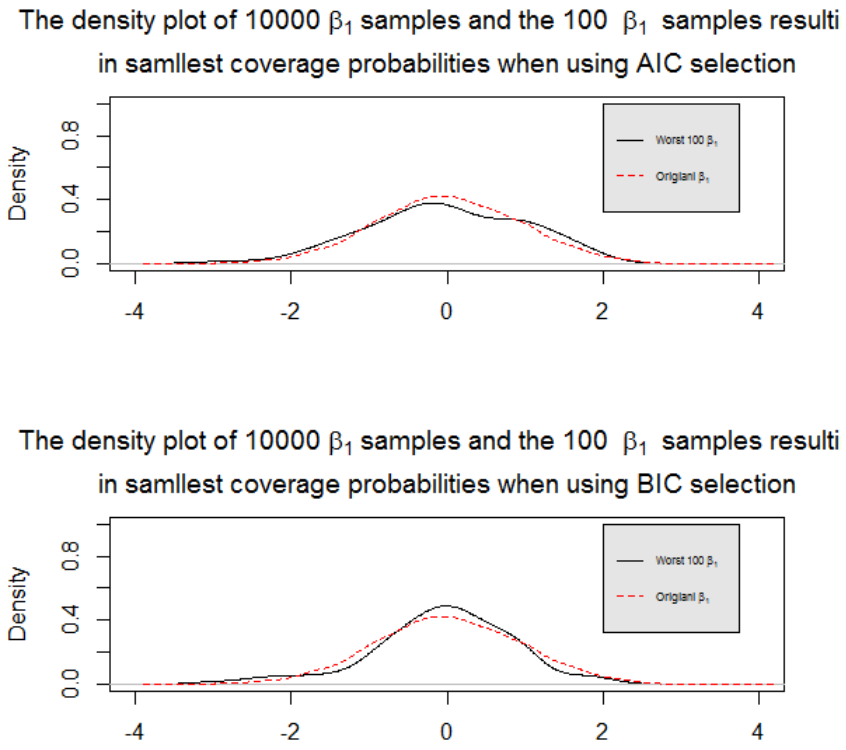

Figure A.2: The density plot of 100 coefficients β_1 which result in 100 smallest coverage probabilities of "naive" intervals for Design 3

References

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate - a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Methodological*, *57*(1), 289-300.

Benjamini, Y., & Yekutieli, D. (2005). False discovery rate–adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, *100*(469), 71-81.

Berk, R., Brown, L., Buja, A., Zhang, K., & Zhao, L. (2013). Valid post-selection inference. *The Annals of Statistics*, *41*(2), 802-837.

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

Fithian, W., Sun, D., & Taylor, J. (2014). Optimal inference after model selection. *arXiv preprint*, *arXiv:1410.2597v2*.

Leeb, H., & Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, *21*(01).

Leeb, H., Pötscher, B. M., & Ewald, K. (2015). On various confidence intervals post-model-selection. *Statistical Science*, *30*(2), 216-227.

Rawlings, J. O., Pantula, S. G., & Dickey, D. A. (1998). *Applied regression analysis : a research tool* (2nd ed.). New York: Springer.

Taylor, J., & Tibshirani, R. J. (2015). Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, *112*(25), 7629-34.