Exact uniformly most powerful post-selection confidence distributions

Andrea C. Garcia-Angulo and Gerda Claeskens
KU Leuven, Belgium

February 2022

Abstract

A conditioning on the event of having selected one model from a set of possibly misspecified normal linear regression models, leads to the construction of uniformly optimal conditional confidence distributions. They can be used for valid post-selection inference. The constructed conditional confidence distributions are finite sample exact and encompass all information regarding the focus parameter in the selected model. This includes the construction of optimal post-selection confidence intervals at all significance levels and uniformly most powerful hypothesis tests.

Keywords: confidence distribution, confidence interval, linear model, model selection, post-selection inference, selective inference, sufficiency.

1 Introduction

We approach the question of valid inference after model selection via confidence distributions and curves. While the well-known confidence intervals are constructed for a single fixed confidence level, a confidence curve may be interpreted as a collection of such confidence intervals for all possible levels. For an overview and historical details about this method see Xie and Singh (2013) and Schweder and Hjort (2016). Cox (1958) introduced the terminology confidence distribution. Using the connection between confidence intervals and hypothesis testing, the confidence curve contains information about the power of the related hypothesis test while p-values are obtained using the cumulative confidence distribution at any hypothesized value for the parameter of interest. Working with confidence distributions provides a more complete picture as opposed to only studying a hypothesis test for a given significance level or a fixed level confidence interval.

Until now, the theory on confidence distributions is based on the assumption that the model is given and correct (Xie and Singh, 2013). In this paper we consider the pre-selection of a model or variables and construct confidence distributions after selection. Hence a main difference between this work and earlier methods is that the model that is used for inference is no longer given beforehand and might be incorrect.

Several classical techniques have been developed to perform model selection. Among the most used ones are information criteria methods (for an overview, see Claeskens and Hjort, 2008), stepwise procedures and within high-dimensional analysis, when the number of covariates p might exceed the number of observations n, regularized estimation procedures such as lasso and least-angle regression (see, e.g., Hastie et al., 2009).

The use of model selection methods is not without a cost. Danilov and Magnus (2004) explicitly warn against overinterpreting results after pretesting has been performed. Kabaila (2009) clearly explains that using classical inference methods after having performed model selection on the same data, can lead to inaccurate conclusions due to confidence regions for model parameters that might have much lower coverage than the nominal coverage value suggests. A same message regarding overoptimistic interpretations by using naive approaches that ignore the selection uncertainty has been told by Hjort and Claeskens (2003), see also Claeskens and Hjort (2008, Chap. 7) where a simulation approach was suggested to construct better confidence intervals post-selection. Kabaila et al. (2016) study the coverage and scaled expected length of confidence intervals for model averaging procedures, a concept that is related to model selection. Hong et al. (2018) provide an explanation explaining why in linear models the confidence intervals have too low coverage by studying the estimated error variance in selected models that contain more parameters than strictly necessary.

There has been a big progress on the topic of post-selection inference. Most of the existing proposals to properly handle this issue can be classified into two groups: a simultaneous inference approach and a conditional approach. The former, proposed by Berk et al. (2013) aims to provide valid inference without restricting to any specific selection method. An advantage of this approach is that the results are valid even when the selection was based upon a graphical exploration of the data. However, the price to pay is that the confidence intervals are relatively wide. Berk et al. (2013) developed a method valid for linear regression models based on the assumption that the true distribution is Gaussian and homoscedastic. For extensions to some other models and accounting for misspecification, see Bachoc et al. (2020). On the other hand, the conditional approach aims to provide valid inference by conditioning the distribution of the parameter estimator of interest by the information that a certain model has been selected. To specify the event of selection one can make use of a so-called selection region which is determined by the specific selection method, which may be comprised of a combination of classical selection methods. This conditional approach is expected to provide narrower confidence intervals than the simultaneous one. Charkhi and Claeskens (2018) developed a conditional approach that provides valid confidence intervals for parameters in likelihood models after model selection by Akaike's information criterion (Akaike, 1973). In selective inference, Lee et al. (2016) provided a method for inference when lasso is used for estimation and selection in linear regression models. Extensions to other selection methods include forward stepwise regression, least angle regression (Tibshirani et al., 2016), marginal screening (Lee and Taylor, 2014) and likelihood and test-based methods (Rügamer and Greven, 2018). Tian and Taylor (2017), Tibshirani et al. (2018) and Taylor and Tibshirani (2018) discuss extensions to non-Gaussian data within affine-selection and Tian and Taylor (2018) provided a more powerful method by the use of randomization. These authors have focused on providing valid p-values and confidence intervals when the significance level is set beforehand. With the use of confidence distributions one obtains information on all significance levels at once. Plots of confidence curves for different methods in a single graph allow for an easy visual comparison of the methods.

Especially when several models are at play it is important to state clearly what the target of inference is. In selective inference this target is a parameter that is present in the selected model. One part of the statistical challenge for correct inference arises due to the fact that the target is only determined after the selection took place. Another difficulty is due to a possible model misspecification. In contrast, when the target is a parameter of the full model, regardless of any selection, inference tools for the possibly misspecified full model can be used. In this paper we address the scenario of selective inference for a focus parameter specified in a selected model.

In Section 2 we describe the modeling framework and review the definition of confidence distributions for a given model. Model selection and the corresponding selection regions are introduced in Section 3. We collect the main theoretical results in Section 4. Computational aspects are discussed in Section 5. Section 6 contains a simulation comparison of the proposed method with some of the methods mentioned above. For instance the simulation results showcase how most of the methods proposed for valid post-selection inference are conservative. A data example is included in Section 7 and Section 8 concludes.

2 Framework, definitions and notation

2.1 Model specification

Let $\boldsymbol{Y}_n = (Y_1, \ldots, Y_n)^\top$ be a n-dimensional vector of independent random variables generated from a multivariate normal distribution $\boldsymbol{Y}_n \sim N_n(\boldsymbol{\mu}, \sigma^2 I_n)$. In this saturated generating distribution, the mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top$ is not further specified. Heteroscedastic errors can be dealt with, see Section 5.

The mean is *modelled* in a linear way using p-vectors of covariates. Let $X = (x_1^\top, \ldots, x_n^\top)^\top$ be a $n \times p$ full rank matrix of fixed or random regressors and denote $\beta = (\beta_1, \ldots, \beta_p)^\top$ a p-dimensional vector of regression parameters. The linear regression model then writes $E(Y_i|x_i) = x_i^\top \beta$ for $i = 1, \ldots, n$. We do *not* assume that such a linear structure is true. For more explanation about the misspecification, see Section 3. In the case of a random design we assume that $(Y_i, x_i^\top)$, $i = 1, \ldots, n$ are independent and identically distributed. For a fixed design the responses remain independent though their mean depends on the covariate value. Throughout the paper we condition on X in all expressions. To not overload the notation, this is not always explicitly indicated.

The normal working density can be rewritten in what is called the natural parametrization of exponential family distributions. The vector of natural parameters for the normal model is $\pi(\beta, \sigma) = (\beta^\top/\sigma^2, -1/(2\sigma^2))^\top$, with the corresponding vector of sufficient statistics $\widetilde{T}(\boldsymbol{Y}_n; X) = (x_1^\top \boldsymbol{Y}_n, \ldots, x_p^\top \boldsymbol{Y}_n, \boldsymbol{Y}_n^\top \boldsymbol{Y}_n)^\top$. With $\kappa(\pi(\beta, \sigma)) = \sum_{i=1}^n (x_i^\top \beta)^2/(2\sigma^2) + n/2 \log(2\pi\sigma^2)$, the normal working density is thus

$$f_n(\boldsymbol{y}_n | X, \pi(\beta, \sigma)) = \exp\left\{\pi(\beta, \sigma)^\top \widetilde{T}(\boldsymbol{y}_n; X) - \kappa(\pi(\beta, \sigma))\right\}. \tag{1}$$

We first single out one regression coefficient for inference. Linear combinations are dealt with in Section 4.2. After a possible reordering, we denote this focus coefficient by θ and combine all other parameters, including the variance parameter in a vector of nuisance parameters η. Thus we redefine $\pi(\beta, \sigma) = (\theta, \eta^\top)^\top$ and we reorder, as elsewhere conditional on X, the vector of sufficient statistics $\widetilde{T}(\boldsymbol{Y}_n; X) = (T(\boldsymbol{Y}_n; X), U^\top(\boldsymbol{Y}_n; X))^\top$, where $T = T(\boldsymbol{Y}_n; X)$ is the sufficient statistic for the scalar parameter of interest θ, and $U(\boldsymbol{Y}_n; X)$ is the vector of sufficient statistics for the nuisance parameters η. For example, we are interested in inference for β_1 and take $\theta = \beta_1/\sigma^2$. Once we condition T on $U(\boldsymbol{Y}_n; X)$, the conditional distribution of $T | U(\boldsymbol{Y}_n; X)$ contains no information about σ^2 as one element of $U(\boldsymbol{Y}_n; X)$, namely $\sum_{i=1}^n Y_i^2$ is sufficient for the single parameter σ^2 up to a known constant.

In the context of model selection when there is more than one model at play, see Section 3, the notation may indicate the specific model as a subscript. For example, in the linear model with mean $X_M \beta_M$ and the natural parameters ordered as $(\theta_M, \eta_M^\top)^\top$, the vector $T_M(\boldsymbol{Y}_n; X_M)$ indicates the sufficient statistic for θ_M and $U_M(\boldsymbol{Y}_n; X_M)$ is the vector of sufficient statistics for the nuisance parameters in this model. Model M's density may also be indicated by $f_{n,M}$ to avoid possible confusion.

2.2 Review of confidence distributions for a parameter - classic case of a given model

A confidence distribution is an inferential tool which summarizes all the information about the parameters of interest carried in the data. As discussed by Singh et al. (2005), Xie and Singh (2013) and Schweder and Hjort (2002, 2016), confidence distributions provide a complete picture for frequentist inference in terms of p-values, confidence intervals and point estimators. We revise first the definition in the case of a correctly specified model, see Schweder and Hjort (2002). See Singh et al. (2005) for an extension to the asymptotic case. In Section 4 we extend this framework to be valid after selection.

Definition 1 *Let $\mathcal{Y}$ be the sample space for the sample data $\boldsymbol{Y}_n = (Y_1, \ldots, Y_n)^\top$ from a parametric distribution F_{θ_0, η_0} where η_0 is the true vector of nuisance parameters and θ_0 is the true value of the parameter of interest with $\Theta \subseteq \mathbb{R}$ as its parameter space. A function $C_n : \Theta \times \mathcal{Y} \to [0, 1]$ is a confidence distribution for $\theta \in \Theta$ if it satisfies:*

(R1) For each given $\boldsymbol{Y}_n = \boldsymbol{y}_n \in \mathcal{Y}$, the function $\theta \mapsto C_n(\theta; \boldsymbol{y}_n)$ is a cumulative distribution function on Θ.

(R2) As a function of $\boldsymbol{Y}_n$, $C_n(\theta_0, \boldsymbol{Y}_n)$ follows a uniform distribution $\mathcal{U}[0, 1]$, where θ_0 is the true value of θ.

Condition (R2) states that at the true parameter value $\theta = \theta_0$ the confidence distribution has a uniform distribution under F_{θ_0, η_0} whatever the true value (θ_0, η_0). This requirement is crucial because it ensures, for instance, that for $\alpha \in (0, 1)$ the coverage probability of a $100(1-\alpha)\%$ confidence interval equals $(1 - \alpha)$ and that the size of a hypothesis test is correct.

The quantiles $q_{\alpha/2} = \{\theta \in \Theta : C_n(\theta, \boldsymbol{Y}_n) = \alpha/2\}$ and $q_{1-\alpha/2} = \{\theta \in \Theta : C_n(\theta, \boldsymbol{Y}_n) = 1 - \alpha/2\}$ are the endpoints of the two-sided $100(1 - \alpha)\%$ confidence intervals. The confidence curve (cc) is defined as follows.

Definition 2 *A confidence curve is the following function of a confidence distribution $C_n(\theta, \boldsymbol{Y}_n)$,*

$$cc_n : \Theta \to [0,1] : \theta \mapsto cc_n(\theta) = |1 - 2C_n(\theta, \boldsymbol{Y}_n)| = \begin{cases} 1 - 2C_n(\theta, \boldsymbol{Y}_n) & \text{if } \theta \le \hat{\theta}_{0.5} \\ 2C_n(\theta, \boldsymbol{Y}_n) - 1 & \text{if } \theta \ge \hat{\theta}_{0.5} \end{cases}$$

where $\hat{\theta}_{0.5} = C_n^{-1}(\frac{1}{2})$ is the median of confidence distribution and can be used as a point estimator.

Confidence distributions can be obtained using the information contained in the likelihood or via a pivotal quantity when it is available. For instance, for any pivot, $\mathrm{piv}(\boldsymbol{Y}_n, \theta)$, we can define $G(u) = P(\mathrm{piv}(\boldsymbol{Y}_n, \theta) \le u)$, then $C_n(\theta, \boldsymbol{Y}_n) = G(\mathrm{piv}(\boldsymbol{Y}_n, \theta))$ if the pivot is increasing in θ (Schweder and Hjort, 2002) and $C_n(\theta, \boldsymbol{Y}_n) = 1 - G(\mathrm{piv}(\boldsymbol{Y}_n, \theta))$ if the pivot is decreasing in θ.

When F_{θ_0, η_0} is an exponential family distribution in its natural parametrization, we can use the information contained in the likelihood and construct $C_n(\theta, \boldsymbol{Y}_n)$ using the sufficient statistic for θ_0 conditioning on the sufficient statistics for η_0. In normal linear regression models $Y = X\beta + \varepsilon$ with $\beta = (\beta_1, \ldots, \beta_p)^\top$ and $\varepsilon \sim N(0, I_n)$ with n the sample size, the confidence distribution for a regression parameter β_r $(r = 1, \ldots, p)$ can be obtained using the t-statistic $\mathcal{T}_r = (\hat{\beta}_r - \beta_r)/\hat{\sigma}_r$ where $\hat{\beta}$ and $\hat{\sigma}^2$ are the maximum likelihood estimators of β and σ^2, $\hat{\sigma}_r^2$ is the (r,r)th diagonal element of $\hat{\Sigma} = \hat{\sigma}^2(X^\top X)^{-1}$. In this case $C_n(\beta_r, \boldsymbol{Y}_n) = F_{n-p}(\mathcal{T}_r)$ where F_{n-p} is the cumulative distribution function of a t-distribution with $n-p$ degrees of freedom (Schweder and Hjort, 2002). The confidence distribution can also be constructed as $C_n(\beta_r, \boldsymbol{Y}_n) = P(T > t_{\mathrm{obs}} \mid U = u_{\mathrm{obs}})$ with T the sufficient statistic for $\beta_r\sigma^{-2}$ with observed value t_{obs} and U the sufficient statistic vector for the nuisance parameters $(\beta_{-r}^\top \sigma^{-2}, -1/2\sigma^{-2})^\top$ with observed value u_{obs} and where the vector β_{-r} denotes the subvector of β that omits the rth component. In this example, both constructions are connected as the maximum likelihood estimator is a function of the sufficient statistics.

It is important to remark that confidence distributions depend on the model and as mentioned by Xie and Singh (2013) the theory on confidence distributions developed until now assumes a correct and pre-specified model. When the working model has been selected using data-driven methods on $\boldsymbol{Y}_n$ and the same data are used for producing the confidence distribution, for example $C_n(\beta_r, \boldsymbol{Y}_n) = F_{n-p}(\mathcal{T}_r)$ or $C_n(\beta_r, \boldsymbol{Y}_n) = P(T > t_{\mathrm{obs}} \mid U = u_{\mathrm{obs}})$ after selection might not longer satisfy (R2) in Definition 1 as we can we see in the next example.

Example. Naive confidence distributions after selection. Figure 1 summarizes the empirical evidence of a small simulation study for inference after variable selection in a linear model. For $i = 1, \ldots, n = 100$, the covariates are independently generated as $x_{i,1} = 1$ and $(x_{i,2}, \ldots, x_{i,6})^\top \sim N(\mathbf{0}_5, \Omega)$ where Ω is the variance-covariance matrix with diagonal entries equal to 1 and off-diagonal elements equal to 0.25. The true values for the parameters are $\beta = (2, -1.5, 0.8, -0.02, 0, 0)^\top$ and $\sigma^2 = 1$. Akaike's information criterion (AIC) (Akaike, 1973, see also Section 3) is used to select a best model from the set of 32 models consisting of all possible submodels of the largest model which has all six covariates; all models include the intercept. The error variance σ^2 is unknown but consistently estimated. In the simulation study we took the case of an overparametrized selected model with parameters $(\beta_1, \ldots, \beta_5)^\top$.

With $\mathcal{T}_r$ the t-statistic with $n-p$ degrees of freedom, we use $C_n(\beta_r, \boldsymbol{Y}_n) = F_{n-p}(\mathcal{T}_r)$ for the selected model and ignore that the variable selection took place. We obtain the cumulative confidence distributions for 1000 samples for which this model is selected.

The drastic effect of selection in naive inference can be clearly observed from the left panel of Figure 1. While naive inference for the parameters β_2 and β_3 is alright, the simulated coverage of the confidence intervals for the truly zero parameter β_5 and for β_4, which has a relatively small value indicates a failure of naive inference. In this case the t-distribution with 95=100-5 degrees of freedom was used to produce confidence intervals for the model parameters. Strikingly, the coverage is zero for all $1 - \alpha$ confidence intervals up to about 0.8 as confidence level. On the q-q plot at the right we can clearly observe that the simulated distributions of the confidence distributions for β_4 and β_5 evaluated at their true value are highly non-uniform, indicating problems with naive inference ignoring the selection and a violation of condition (R2) in Definition 1.

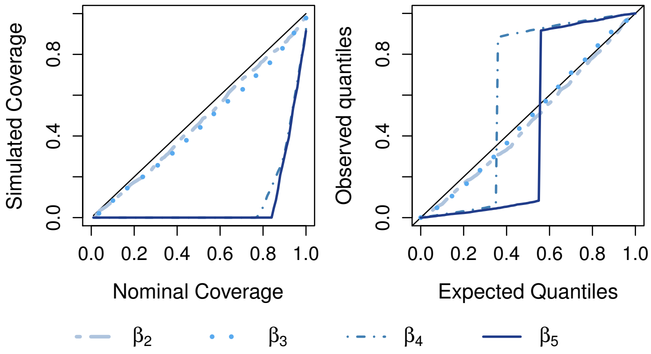

Figure 1: Left: Mean simulated coverage of $1-\alpha$ confidence intervals with $\alpha = [0,1]$ for the regression parameters $\beta_2, \ldots, \beta_5$ in the selected linear model for a naive method that ignores selection. Each line corresponds to one of the four regression parameters subject to selection. Right: Quantiles of the simulated distribution of $C_n(\beta_{0,r}, \boldsymbol{Y}_n)$ versus expected quantiles of a $\mathcal{U}[0,1]$, showing a clear deviation from uniformity and failure of naive inference for β_4 and β_5.

3 Model selection methods and regions

We perform selection within a model selection set $\mathcal{M} = \{M_1, \ldots, M_m\}$ with a finite number, m, candidate normal linear models. Each model $M \in \mathcal{M}$ uses its own full rank $n \times p_M$ design matrix X_M to specify $E[\boldsymbol{Y}_n \mid X_M] = X_M \beta_M$. Thus p_M denotes the number of regression coefficients for model M such that β_M is a vector of length p_M. The models in $\mathcal{M}$ do not need to be nested and we explicitly do *not* assume the linear structure to be true.

Since the true data generating density corresponds to a saturated model, all linear models in $\mathcal{M}$ are possibly misspecified. The estimator of β_M targets the pseudo-true parameter β_M^*, that is, the parameter value which minimizes the Kullback-Leibler distance between the model density and the true data generating density (White, 1994). For the normal model with mean vector μ, the target is explicitly obtained as $\beta_M^* = (X_M^\top X_M)^{-1} X_M^\top \mu$, see also Lee et al. (2016).

Each model selection method comes with its own partitioning of the data sample space $\mathcal{Y} = \cup_{j=1}^m A_j$ with $A_k \cap A_l = \emptyset$ if $k \neq l$ such that the result of selecting model M_j is equivalent with $\boldsymbol{Y}_n \in A_j$. Indeed, model selection corresponds to picking a single model as the selected one. Charkhi and Claeskens (2018) find regions for AIC selection characterized via inequalities using quadratic forms. Selection regions corresponding to model selection based on likelihood ratio testing, F-tests or more general 'significance hunting' by t-tests are described by Rügamer and Greven (2018). Lee et al. (2016) obtain polyhedral regions for lasso selection, while Loftus and Taylor (2015) characterizes selection via cross-validation.

We assume that all models in $\mathcal{M}$ have a nonzero selection probability and that the selection regions are expressible in terms of the sufficient statistics for the model parameters. We state some examples.

Examples. *(1) Selection by AIC or BIC.* Penalized likelihood-based information criteria such as AIC (Akaike, 1973) which uses $\text{pen}_{M_j} = 2|M_j|$, with $|M_j|$ the number of estimated parameters in model M_j and BIC (Schwarz, 1978) with $\text{pen}_{M_j} = \log(n)|M_j|$ attach a value to each model. The model with the lowest such value is selected. For selecting model $M_j \in \mathcal{M}$ with $j \in \{1, \ldots, m\}$ the selection region which indicates for which sample values this model gets selected is obtained as

$$\begin{aligned} A_j \quad = \quad & \{\boldsymbol{y}_n \in \mathbb{R}^n : -2\log f_{n,M_j}(\boldsymbol{y}_n | X_{M_j}, \pi(\hat\beta_{M_j}, \hat\sigma_{M_j})) + \text{pen}_{M_j} \\ & < -2\log f_{n,M_k}(\boldsymbol{y}_n | X_{M_k}, \pi(\hat\beta_{M_k}, \hat\sigma_{M_k})) + \text{pen}_{M_k}, \text{for all } M_k \in \mathcal{M} \setminus M_j\}. \end{aligned}$$

Maximum likelihood estimators are used to estimate each model's parameters. The inequalities defining A_j can equivalently be expressed in terms of sufficient statistics. Indeed, since maximum likelihood estimators are functions of the sufficient statistics too, $A_j = \{\boldsymbol{y}_n \in \mathbb{R}^n : 2\pi(\hat\beta_{M_k}, \hat\sigma_{M_k})^\top \widetilde{T}_{M_k}(\boldsymbol{y}_n; X_{M_k}) - 2\pi(\hat\beta_{M_j}, \hat\sigma_{M_j})^\top \widetilde{T}_{M_j}(\boldsymbol{y}_n; X_{M_j}) + 2\kappa(\pi(\hat\beta_{M_j}, \hat\sigma_{M_j})) - 2\kappa(\pi(\hat\beta_{M_k}, \hat\sigma_{M_k})) + \text{pen}_{M_j} - \text{pen}_{M_k} < 0, \text{for all } M_k \in \mathcal{M} \setminus M_j\}$.

In a similar way one obtains selection regions in terms of sufficient statistics when a likelihood ratio test determines the best model in a set of nested models.

(2) Backward selection via t-tests. Another commonly used selection procedure, even despite the criticism it received, is backward variable selection by using the significance of t-tests. This procedure is an example of "significance hunting" described in Berk et al. (2013) and Rügamer and Greven (2018). The method starts with a full model M_1 with p parameters. At each step $s = 1, \ldots, p$, there are $p - s + 1$ null hypotheses tested $H_{0,r} : \beta^*_{M,r} = 0$ for $r = 1, \ldots, p - s + 1$ at significance level α and the parameter with the largest p-value indicating insignificance is discarded. Define $\mathcal{T}_{M,r} = \hat{\Sigma}^{-1/2}_{M,r}\ \hat{\beta}_{M,r}$, the t-statistic for the rth component of the estimated parameter vector in model M, $I(\cdot)$ the indicator function and $t_{\alpha/2}$ the critical value given by the $(1-\alpha/2)$-quantile of a Student's t-distribution with $n - (p - s + 2)$ degrees of freedom. Both the estimator $\hat{\beta}_{M,r}$ and its estimated variance $\hat{\Sigma}_{M,r}$ can be expressed in terms of sufficient statistics. Thus $\beta_{M,d}$ is omitted from the model when $d = \arg\min_r\{|\mathcal{T}_{M,r}|\ I(|\mathcal{T}_{M,r}| \le t_{\alpha/2})\}$. This testing procedure is repeated until all indicators equal 0 leading to the selected model $M_{\hat{\jmath}}$. Note that $\mathcal{M}$ is not fixed beforehand but it is updated at each step. An illustration of the procedure is as follows. Starting with the full model M_1, parameter $\beta_{[1]}$ is discarded from the model when

$$\bigcap_{r\in\{1,\ldots,p\}\setminus\{[1]\}} \Big\{|\mathcal{T}_{M_1,[1]}| \le |\mathcal{T}_{M_1,r}|\Big\} \cap \big\{|\mathcal{T}_{M_1,[1]}| \le t_{\alpha/2}\big\}\,.$$

The model without $\beta_{[1]}$ is denoted by M_2. A variable $\beta_{[2]}$ is omitted from M_2 when

$$\bigcap_{r\in\{1,\ldots,p-1\}\setminus\{[2]\}} \Big\{|\mathcal{T}_{M_2,[2]}| \le |\mathcal{T}_{M_2,r}|\Big\} \cap \big\{|\mathcal{T}_{M_2,[2]}| \le t_{\alpha/2}\big\}\,,$$

etc. until no variables are insignificant. The combination of such events defines the selection regions in terms of the sufficient statistics $\widetilde{T}_M(\boldsymbol{y}_n; X_M)$ in each model $M \in \{M_1, \ldots, M_p\}$. □

4 Optimality results of post-selection conditional confidence distributions

A model selection procedure is carried out and one observes which parameters appear in the selected model. At that moment one decides to perform inference for one of these parameters, say θ. In this case the framework of selective inference applies.

For a given sample $\boldsymbol{Y}_n = \boldsymbol{y}_n$, we denote the selected model by $M_{\hat{\jmath}}$, which is equivalent to stating that $\boldsymbol{y}_n \in A_{\hat{\jmath}}$. The notation with the hat on the subscript j should remind us that the model was selected based on the sample. Let θ be the parameter of interest in this model with parameter space Θ. Conditional on the selection and under a possibly misspecified model $M_{\hat{\jmath}}$, we define the **conditional post-selection confidence distribution** as follows.

Definition 3 *Let $\mathcal{M} = \{M_1, \ldots, M_m\}$ be a model selection set to which a selection criterion with sample space partitioning $\mathcal{Y} = \cup_{j=1}^m A_j$ is applied. Conditional on the selected model $M_{\hat{\jmath}}$ with pseudo true-parameter vector $(\theta^*_{M_{\hat{\jmath}}}, \eta^{*\top}_{M_{\hat{\jmath}}})$, a function $C_{n|\hat{\jmath}} : \Theta \times A_{\hat{\jmath}} \to [0,1] : (\theta, \boldsymbol{Y}_n) \mapsto C_{n|\hat{\jmath}}(\theta, \boldsymbol{Y}_n)$ is a conditional post-selection confidence distribution if it satisfies:*

(R3) For each given $\boldsymbol{Y}_n = \boldsymbol{y}_n \in A_{\hat{\jmath}}$, the function $\Theta \to [0,1] : \theta \mapsto C_{n|\hat{\jmath}}(\theta, \boldsymbol{y}_n)$ is a cumulative distribution function on Θ.

(R4) As a function of $\boldsymbol{Y}_n \in A_{\hat{\jmath}}$, $C_{n|\hat{\jmath}}(\theta^*_{M_{\hat{\jmath}}}, \boldsymbol{Y}_n)$ follows a uniform distribution $\mathcal{U}[0,1]$ whatever the value of the pseudo true parameter vector for this model.

Note that the subscript '$|\hat{\jmath}$' reminds about the conditioning on the selection event. Conditioning on the selection is needed to avoid that the function domain changes for each sample. When $M_{\hat{\jmath}}$ is the selected model we only consider the samples in $A_{\hat{\jmath}}$, thus for which the selection leads to the same selected model $M_{\hat{\jmath}}$. By working with the pseudo-true parameters, we take into account that the models do not need to be correctly specified. Thus (R4) guarantees the correct coverage of the confidence curves as in (R2) but with respect to the pseudo true-value $\theta^*_{M_{\hat{\jmath}}}$.

A conditional post-selection confidence curve is defined in analogy to Definition 2 as a function from $\Theta \to [0,1]$ such that $\theta \mapsto cc_{n|\hat{\jmath}}(\theta) = |1 - 2C_{n|\hat{\jmath}}(\theta, \boldsymbol{Y}_n)|$. For example, for $\alpha = 0.95$, the distinct values $\theta_1 < \theta_2$ such that $cc_{n|\hat{\jmath}}(\theta_1) = cc_{n|\hat{\jmath}}(\theta_2) = 0.95$ form the endpoints of the 95% confidence interval $[\theta_1, \theta_2]$ for $\theta^*_{M_{\hat{\jmath}}}$. One strong advantage of working with a confidence curve is that it contains all information regarding confidence and is not restricted to a single pre-specified confidence level.

4.1 Uniformly most powerful confidence distributions

Since the selection of a model $M_{\hat{\jmath}}$ depends on the set of models $\mathcal{M}$, this information is taken into account by conditioning the sufficient statistic $T = T(\boldsymbol{Y}_n; X_{M_{\hat{\jmath}}})$ for the focus parameter θ on a vector of sufficient statistics for all nuisance parameters. We define $U_{\mathcal{M}} = U_{\mathcal{M}}(\boldsymbol{Y}_n)$ the vector with components in some arbitrary but fixed order from the collection $\{U_M(\boldsymbol{Y}_n; X_M)$ for all $M \in \mathcal{M}\}$ where duplicated items are removed. We use the distribution of T conditional on both $U_{\mathcal{M}}$ and the selection event $\boldsymbol{Y}_n \in A_{\hat{\jmath}}$ to obtain the conditional post-selection confidence distribution. After conditioning the domain of T is restricted to $\text{dom}(T|U_{\mathcal{M}} = u_{\text{obs}}, \boldsymbol{Y}_n \in A_{\hat{\jmath}}) = \{\boldsymbol{y}_n \in A_{\hat{\jmath}} : U_{\mathcal{M}}(\boldsymbol{y}_n) = u_{\text{obs}}\}$, which is fixed. Let $\boldsymbol{y}_n$ denote the observed value of $\boldsymbol{Y}_n$. We denote by $u_{\text{obs}} = U_{\mathcal{M}}(\boldsymbol{y}_n)$ the vector that consists of the observed values of the sufficient statistics for the nuisance parameters for all models in $\mathcal{M}$. For normal linear regression models this domain can be exactly computed. For some examples, see Rügamer and Greven (2018) when the selection is by likelihood based model selection procedures for Gaussian data and explicit derivations are obtained in simple two-model comparisons.

The optimality properties for hypothesis testing in an exponential family have been widely studied (Lehmann and Scheffé, 1955; Lehmann and Romano, 2006) and such properties can be extended to confidence distributions. Optimality for a confidence distribution is expressed in terms of confidence loss, $\text{loss}(\theta_a, C) = \int B(\theta_a^{'} - \theta_a)\; dC(\theta_a^{'}, Y)$ associated with a function B, nondecreasing on the positive half-axis, nonincreasing on the negative half-axis and $B(0) = 0$. For instance, $B(x) = x^2$ is the quadratic function for which $\text{loss}(\theta_a, C)$ is the squared loss. A confidence distribution is uniformly optimal if (Schweder and Hjort, 2016, Def 5.9) for every B, defined as above, $\text{loss}(\theta_a, C_{opt}) \leq \text{loss}(\theta_a, C)$ for any other C, for every value θ_a.

The propositions below are an extension of Theorem 5.11 of Schweder and Hjort (2016) for the case of post-selection inference. In proposition 1, we first work under the assumption that σ^2 is known or estimated independently of the data $\boldsymbol{Y}_n$ used for inference. We relax that assumption in proposition 2 at the cost of a extra conditioning. Note that these are a exact finite sample results, no asymptotic statements are involved. The proofs are contained in the Appendix.

Proposition 1 *Let $M_{\hat{\jmath}}$ be selected from a set of linear models $\mathcal{M}$ for the data $\boldsymbol{Y}_n$ conditional on covariates X. We assume that all models in $\mathcal{M}$ have a nonzero selection probability and that the selection regions can be expressed in terms of the sufficient statistics for the model parameters. We assume σ^2 is known or independently estimated for all models in $\mathcal{M}$. Let $\theta_{M_{\hat{\jmath}}}$ be the univariate parameter of interest in $M_{\hat{\jmath}}$ with parameter space Θ and sufficient statistic $T = T(\boldsymbol{Y}_n; X_{M_{\hat{\jmath}}})$. The corresponding pseudo-true parameter value is denoted by $\theta^*_{M_{\hat{\jmath}}}$. Let $U_{\mathcal{M}} = U_{\mathcal{M}}(\boldsymbol{Y}_n)$ be the combined vector of sufficient statistics for the nuisance parameters η^* in $\mathcal{M}$. The observed values of T and $U_{\mathcal{M}}$ are denoted by t_{obs} and u_{obs}. The conditional post-selection confidence distribution:*

$$C_{n,|\hat{\jmath}} : \Theta \times A_{\hat{\jmath}} \to [0,1] : (\theta, \boldsymbol{Y}_n) \mapsto P(T > t_{\text{obs}} \mid U_{\mathcal{M}} = u_{\text{obs}}, \boldsymbol{Y}_n \in A_{\hat{\jmath}}) \tag{2}$$

is the uniformly most powerful (UMP) conditional post-selection confidence distribution for $\theta^*_{M_{\hat{\jmath}}}$.

When σ^2 is unknown, which is more common in practice, an additional conditioning is required. Define the projection matrix in model M by $P_{X_M} = X_M(X_M^\top X_M)^{-1}X_M^\top$ and define for the selected model $M_{\hat{\jmath}}$ the statistic $V = V(\boldsymbol{Y}_n) = (I - P_{X_{M_{\hat{\jmath}}}})\boldsymbol{Y}_n$. To account for the estimation of σ^2 a conditioning on V is required as stated in Proposition 2. This conditioning is also used in selective inference (see, e.g. Fithian et al., 2017; Tian et al., 2018).

Proposition 2 *Let $M_{\hat{\jmath}}$ be selected from a set of linear models $\mathcal{M}$ for the data $\boldsymbol{Y}_n$ conditional on covariates X. We assume that all models in $\mathcal{M}$ have a nonzero selection probability and that the selection regions can be expressed in terms of the sufficient statistics for the model parameters. Let $\theta^*_{M_{\hat{\jmath}}}$ be the univariate pseudo-true parameter in $M_{\hat{\jmath}}$ that is of interest with parameter space Θ and sufficient statistic $T = T(\boldsymbol{Y}_n; X_{M_{\hat{\jmath}}})$. Let $U_{\mathcal{M}} = U_{\mathcal{M}}(\boldsymbol{Y}_n)$ be the combined vector of sufficient statistics for the nuisance parameters η^* in $\mathcal{M}$ and $V = (I - P_{X_{M_{\hat{\jmath}}}})\boldsymbol{Y}_n$. The observed values of T, $U_{\mathcal{M}}$ and V are denoted by t_{obs}, u_{obs} and v_{obs}. We assume that the $(n+p+1)$-dimensional parameter space for $(\theta^*_{M_{\hat{\jmath}}}, \eta^{*\top}, \mu^\top)^\top$ contains an open rectangle in $\mathbb{R}^{n+p+1}$ and that the sample space does not depend on the parameters. The conditional post-selection confidence distribution:*

$$C_{n,|\hat{\jmath}} : \Theta \times A_{\hat{\jmath}} \to [0,1] : (\theta, \boldsymbol{Y}_n) \mapsto P(T > t_{\text{obs}} \mid U_{\mathcal{M}} = u_{\text{obs}}, V = v_{\text{obs}}, \boldsymbol{Y}_n \in A_{\hat{\jmath}}) \qquad (3)$$

is the uniformly most powerful (UMP) conditional post-selection confidence distribution for $\theta^*_{M_{\hat{\jmath}}}$.

Corollary 1 *Propositions 1 and 2 imply that for a selected model $M_{\hat{\jmath}}$ with pseudo-true parameter of interest $\theta^*_{M_{\hat{\jmath}}}$ conditional on this model being selected:*

(i) For each value $\theta \in \Theta$, $C_{n,|\hat{\jmath}}(\theta, \boldsymbol{y}_n)$ is the p-value of the uniformly most powerful unbiased test for testing $H_0 : \theta^*_{M_{\hat{\jmath}}} = \theta$ against $\theta^*_{M_{\hat{\jmath}}} > \theta$.

(ii) The confidence curve $\Theta \to [0,1] : \theta \mapsto cc_{n,|\hat{\jmath}}(\theta)$ provides the shortest $100(1-\alpha)\%$ confidence intervals for the pseudo-true value $\theta^*_{M_{\hat{\jmath}}}$ for every $0 < \alpha < 1$ among all other coverage proper confidence curves.

Another interesting consequence from Propositions 1 and 2 is that we can construct an alternative estimator for the focus parameter using the post-selection confidence distribution $C_{n,|\hat{\jmath}}$. A graphical representation of the confidence curve which shows the confidence interval boundaries versus the confidence level when considered at level zero immediately points towards the median confidence estimator for the focus parameter $\theta^*_{M_{\hat{\jmath}}}$ defined by $\tilde{\theta}_{n,0.5} = C^{-1}_{n,|\hat{\jmath}}(0.5)$. Yet alternative estimators are the mean of the post-selection confidence distribution and the mode of the post-selection confidence density. Xie and Singh (2013) in their theorem 1 show the consistency of these estimators under a given parametric data generating density.

In particular, for the median post-selection confidence estimator we have the following result, adapted from Singh et al. (2007, Theorem 3.1) to the misspecified case using the pseudo-true parameter vector.

Corollary 2 *Under the assumptions of Propositions 1 or 2. For any $\epsilon \in (0, 0.5)$, if $L_n(\epsilon) = C^{-1}_{n,|\hat{\jmath}}(1-\epsilon) - C^{-1}_{n,|\hat{\jmath}}(\epsilon) \to 0$ in probability as $n \to \infty$, then the median post-selection confidence estimator $\tilde{\theta}_{n,0.5}$ is consistent for $\theta^*_{M_{\hat{\jmath}}}$. If additionally $L_n(\epsilon) = O_p(a_n)$ for a non-negative sequence $a_n \to 0$, then $\tilde{\theta}_{n,0.5} - \theta^*_{M_{\hat{\jmath}}} = O_P(a_n)$.*

One crucial assumption regarding the model selection methods is that the selection regions which form a partition of the sample space can be expressed in terms of the sufficient statistics of the model parameters; see Section 3 for some examples. This is the case for selection methods using information criteria that are maximum likelihood-based, such as AIC and BIC. Also the forward and backward search procedures using t-tests, F-tests or likelihood ratio tests are included (Rügamer and Greven, 2018), as is the selection by cross-validation or k-fold cross-validation (see Loftus, 2015). Some methods that cannot be expressed using sufficient statistics of the model parameters include model changes after visual inspections of the data via, for example, residual plots or histograms. Tibshirani et al. (2016) obtain selection regions in the form of polyhedral sets for forward stepwise regression, least angle regression, and the lasso. They used an additional conditioning on the active signs of the estimates in the selected model and mentioned that this is merely out of computational convenience. In Rügamer and Greven (2018) for inference on a linear combination of the form $v^\top \mu$ with μ the unspecified true mean vector there is a conditioning on the quadratic inequality that determines which model is selected as well as on a projection of the response Y on the space orthogonal to the vector v. Our result shows that a conditioning on the sufficient statistics for the nuisance parameters in the complete model set $\mathcal{M}$ leads to optimal results. This agrees with the recommendation of Tibshirani et al. (2016) for the use of forward stepwise regression to obtain the most powerful selective test for a given level α. The conditioning on all sufficient statistics, not only those of the selected model is necessary in order to fix the domain which otherwise might depend on random quantities. Our results are valid for all $\alpha \in [0,1]$ and provide a complete picture for optimal post-selection inference for the focus parameter, both regarding hypothesis testing as well as the construction of confidence intervals at all confidence levels.

4.2 Linear combinations of parameters

As in Section 2.1, we consider a linear model with a p-vector β of regression parameters and with variance σ^2. The $k = (p+1)$-dimensional vector of natural parameters $\pi(\beta,\sigma) = (\beta^\top/\sigma^2, -1/(2\sigma^2))^\top$. Consider now the linear combination $\psi = \sum_{r=1}^J c_r\pi_r(\beta,\sigma)$, where the c_r are given constants and $\pi_r(\beta,\sigma)$ are the elements of $\pi(\beta,\sigma)$ for $r = 1,\ldots,R \leq k$. A typical example is a linear combination of the form $x_0^\top\beta$. Without loss of generalization, let $c_1 \neq 0$ and $R = k$ then

$$\pi_1(\beta,\sigma) = \frac{\psi - \sum_{r=2}^k c_r\pi_r(\beta,\sigma)}{c_1},$$

and we may write

$$\begin{aligned}(\pi(\beta,\sigma)^\top\widetilde{T}(\boldsymbol{y}_n;X)) &= \frac{\psi - \sum_{r=2}^k c_j\pi_r(\beta,\sigma)}{c_1}T_1(\boldsymbol{y}_n;X) + \sum_{r=2}^k \pi_r(\beta,\sigma)T_r(\boldsymbol{y}_n;X)\\ &= \psi\frac{T_1(\boldsymbol{y}_n;X)}{c_1} + \sum_{r=2}^k \pi_r(\beta,\sigma)\left(T_r(\boldsymbol{y}_n;X) - \frac{c_rT_1(\boldsymbol{y}_n;X)}{c_1}\right),\end{aligned}$$

which shows that, under a reparametrization, the distribution in (1) is also of the exponential family form in its natural parametrization with the k-dimensional vector of natural parameters $(\psi, \pi_2(\beta,\sigma),\ldots,\pi_k(\beta,\sigma))^\top$ with sufficient statistics

$$(T_\psi, U_\psi^\top) = \left(\frac{T_1(\boldsymbol{Y}_n;X)}{c_1}, \left(T_2(\boldsymbol{Y}_n;X) - \frac{c_2T_1(\boldsymbol{Y}_n;X)}{c_1}\right),\ldots,\left(T_k(\boldsymbol{Y}_n;X) - \frac{c_kT_1(\boldsymbol{Y}_n;X)}{c_1}\right)\right),$$

with $T_\psi = c_1^{-1}T_1(\boldsymbol{Y}_n;X)$ and U_ψ the vector of sufficient statistics for the nuisance parameters for the linear combination (Young and Smith, 2005, Chap. 7).

Now we consider a set of models $\mathcal{M}$, perform model selection and consider the focus parameter ψ in the selected model $M_{\hat{\jmath}}$. Similar as before, we denote by $U_{\mathcal{M},\psi}$ the combined vector of sufficient statistics for the nuisance parameters appearing in the model selection set $\mathcal{M}$.

To apply our theory to obtain post-selection confidence distributions for ψ, note that under the same assumptions as in Proposition 1 the selection region $A_{\hat{\jmath}}$ is also a function of $(T_\psi, U_{\mathcal{M},\psi}^\top)$ after the reparametrization, which means that after selection the domain of $T_\psi|(U_{\mathcal{M},\psi} = u_{\psi,\text{obs}}, \boldsymbol{Y}_n \in A_{\hat{\jmath}})$ becomes $\text{dom}(T_\psi|U_{\mathcal{M},\psi} = u_{\psi,\text{obs}}, \boldsymbol{Y}_n \in A_{\hat{\jmath}}) = \{\boldsymbol{y}_n \in A_{\hat{\jmath}} : U_{\mathcal{M},\psi}(\boldsymbol{y}_n) = u_{\psi,\text{obs}}\}$. This can be extended to the extra conditioning on $V = v_\text{obs}$ as in Proposition 2.

Corollary 3 *Under the assumptions of Propositions 1 or 2, the confidence distribution for* $\psi^*_{M_{\hat{\jmath}}} = \sum_{r=1}^R c_r\pi_r(\beta^*_{M_{\hat{\jmath}}}, \sigma^*_{M_{\hat{\jmath}}})$ *in the selected model* $M_{\hat{\jmath}}$ *with known or independently estimated variance,*

$$C_{n,|\hat{\jmath}}(\psi,\boldsymbol{Y}_n) = P(T_\psi > t_{\psi,\text{obs}} \mid U_\psi = u_{\psi,\text{obs}}, \boldsymbol{Y}_n \in A_{\hat{\jmath}})$$

or in the selected model $M_{\hat{\jmath}}$ *with estimated variance,*

$$C_{n,|\hat{\jmath}}(\psi,\boldsymbol{Y}_n) = P(T_\psi > t_{\psi,\text{obs}} \mid U_\psi = u_{\psi,\text{obs}}, V = v_\text{obs}, \boldsymbol{Y}_n \in A_{\hat{\jmath}}) \tag{4}$$

are the uniformly most powerful (UMP) conditional post-selection confidence distributions for $\psi^*_{M_{\hat{\jmath}}}$ *for the cases when* σ^2 *is known or unknown.*

Example. Let $M_{\hat{\jmath}}$ be the selected linear normal model $\boldsymbol{Y}_n = X_{M_{\hat{\jmath}}}\beta_{M_{\hat{\jmath}}} + \varepsilon$, with $\varepsilon \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 I_n)$ and $p_{M_{\hat{\jmath}}} = q$ with covariates $x_1, \ldots, x_q$ in the design matrix $X_{M_{\hat{\jmath}}}$. We are interested in inference for $\psi^*_{M_{\hat{\jmath}}} = \sum_{j=1}^q c_j \beta^*_{M_{\hat{\jmath}},j}$, a linear combination of the pseudo-true regression parameters for the selected model. Define $\psi = \sum_{j=1}^q c_j\beta_j/\sigma^2$, then $\beta_1/\sigma^2 = \psi/c_1 - \sum_{j=2}^q c_j\beta_j/(c_1\sigma^2)$ and we can write

$$\pi(\beta,\phi)^\top \widetilde{T}_{M_{\hat{\jmath}}}(\boldsymbol{y}_n; X_{M_{\hat{\jmath}}}) = \frac{-1}{2\sigma^2}\sum_{i=1}^n y_i^2 + \psi\frac{\sum_{i=1}^n x_{1i}y_i}{c_1} + \sum_{r=2}^q \frac{\beta_r}{\sigma^2}\left(\sum_{i=1}^n x_{ri}y_i - c_r\frac{\sum_{i=1}^n x_{1i}y_i}{c_1}\right).$$

We take σ^2 as unknown and estimated using the data. Set $T_{M_{\hat{\jmath}},\psi} = \sum_{i=1}^n x_{1i}y_i/c_1$ and $U_{M_{\hat{\jmath}},\psi} = \left(\sum_{i=1}^n y_i^2, \left(\sum_{i=1}^n x_{2i}y_i - c_2\sum_{i=1}^n x_{1i}y_i/c_1\right), \ldots, \left(\sum_{i=1}^n x_{qi}Y_i - c_q\sum_{i=1}^n x_{1i}y_i/c_1\right)\right)$. Since the selection region can be written in function of the sufficient statistics, the event $\boldsymbol{Y}_n \in A_{\hat{\jmath}}$ implies a truncated domain of $T_\psi|(U_{\mathcal{M},\psi} = u_{\psi,\text{obs}}, V = v_{\text{obs}}, \boldsymbol{Y}_n \in A_{\hat{\jmath}})$ after selection of $M_{\hat{\jmath}}$. We obtain the $C_{n,|\hat{\jmath}}(\psi, \boldsymbol{Y}_n)$ in (4) which is also a UMP conditional post-selection confidence distribution for $\psi^*_{M_{\hat{\jmath}}}$. The post-selection confidence distribution $C_{n,|\hat{\jmath}}(\psi^*, \boldsymbol{Y}_n)$, as the conditional distribution of $T_\psi|(U_{\mathcal{M},\psi} = u_{\text{obs}}, V = v_{\text{obs}})$, contains no information about σ^2 once we have conditioned on $\sum_{i=1}^n Y_i^2$.

5 A Monte-Carlo sampling approach

While for some model selection sets the conditional post-selection confidence distribution might be explicit to get, we here provide a simulation approach in case the exact calculation might be difficult to derive explicitly. We can approximate the confidence distribution by using a Monte-Carlo resampling method conditional on the sufficient statistics (Lindqvist and Taraldsen, 2005) as explained by Schweder and Hjort (2016, Ch. 8) with, however, an additional constraint given by the selection event $\boldsymbol{Y}_n \in A_{\hat{\jmath}}$, expressed in terms of the sufficient statistics, see Section 3.Let $M_{\hat{\jmath}}$ be the selected model. The procedure is as follows:

1. For the observed sample $(\boldsymbol{y}_n, X_{M_{\hat{\jmath}}})$ calculate the observed values of the sufficient statistics in the selected model: $t_{\text{obs}} = T_{M_{\hat{\jmath}}}(\boldsymbol{y}_n; X_{M_{\hat{\jmath}}})$, $u_{\mathcal{M},\text{obs}} = U_{\mathcal{M}}(\boldsymbol{y}_n)$ and $v_{\text{obs}} = V(\boldsymbol{y}_n)$.

2. We choose a set of candidate values for the parameter of interest $\theta^*_{M_{\hat{\jmath}}}$.

3. For each candidate parameter value ϑ a big enough number, say B, of samples $\boldsymbol{y}_{n,b} = (y_{1,b}, \ldots, y_{n,b})$ for $b = 1, \ldots, B$, are generated with density $f_{n,M_{\hat{\jmath}}}(\boldsymbol{y}|X, \vartheta, \eta^o_{M_{\hat{\jmath}}})$ which specifies the parameters in the original density $f_{n,M_{\hat{\jmath}}}(\boldsymbol{y}|X, \theta_{M_{\hat{\jmath}}}, \eta_{M_{\hat{\jmath}}})$ by the values of the candidate ϑ and the vector $\eta^o_{M_{\hat{\jmath}}}$ such that the following constraints hold: $U_{\mathcal{M}}(\boldsymbol{y}_{n,b})$ is equal to $u_{\mathcal{M},\text{obs}}$, $V(\boldsymbol{y}_{n,b})$ is equal to v_{obs} and $\boldsymbol{y}_{n,b} \in A_{\hat{\jmath}}$. The constraint $V(\boldsymbol{y}_{n,b}) = v_{\text{obs}}$ is only necessary when σ^2 is unknown and estimated using the observed data, otherwise, it can be omitted.

4. For each generated sample $\boldsymbol{y}_{n,b}$, calculate $T_{M_{\hat{\jmath}}}(\boldsymbol{y}_{n,b}; X_{M_{\hat{\jmath}}})$ and obtain $r(\vartheta) = \sum_{b=1}^B I\{T_{M_{\hat{\jmath}}}(\boldsymbol{y}_{n,b}; X_{M_{\hat{\jmath}}}) > t_{\text{obs}}\}/B$, which is the simulated cumulative confidence distribution at value ϑ.

5. We obtain an approximation of the function $\Theta \to [0,1] : \theta \mapsto C_{n,|\hat{\jmath}}(\theta, \boldsymbol{y}_n)$ by a linear interpolation of the set of points $(\vartheta, r(\vartheta))$.

Due to the simulation variability, this approximation may lead to simulated distributions that are not exactly monotone, which might be remedied by smoothing the approximate confidence distribution under monotonicity constraints (see, e.g. Pya and Wood, 2015).

The constrained sample generation in the second step can be done in two steps. First we generate under the constraints $U_{\mathcal{M}}(\boldsymbol{y}_{n,b}) = u_{\mathcal{M},\text{obs}}$ and $V(\boldsymbol{y}_{n,b}) = v_{\text{obs}}$ which can be done by fine-tuning η until the equalities hold, see for example Schweder and Hjort (2016, Section 8.5) or we can transform it into an optimization problem in which we minimize the squared difference between the observed $u_{\mathcal{M},\text{obs}}, v_{\mathcal{M},\text{obs}}$ and the values $U_{\mathcal{M}}(\boldsymbol{y}_{n,b}), V(\boldsymbol{y}_{n,b})$ computed by the simulated value, where the minimization is over the nuisance parameters η as to satisfy a virtually zero tolerance for the minimized value.

In a second step, we check whether the generated sample which already satisfies $U_{\mathcal{M}}(\boldsymbol{y}_{n,b}) = u_{\mathcal{M},\text{obs}}$ and $V(\boldsymbol{y}_{n,b}) = v_{\text{obs}}$ also belongs to the selection region $A_{\hat{\jmath}}$ by applying the selection procedure to that generated sample. We redo this until we have B samples satisfying all constraints.

In case of misspecification due to heteroscedasticity, the distribution of T conditional on $U_{\mathcal{M}} = u_{\mathcal{M},\text{obs}}$ and $V = v_{\text{obs}}$ has more variability than accounted for by the sufficient statistic of the scale parameter in the selected model. In this case, we might increase the variability in step 3 above by generating B times ϑ_b from $N(\vartheta, \widetilde{\sigma}^2(\theta_{M_{\hat{\jmath}}}))$ for each candidate value ϑ. Here, $\widetilde{\sigma}(\theta_{M_{\hat{\jmath}}})$ is an estimate of the standard error of $\hat{\theta}_{M_{\hat{\jmath}}}$ using a heteroscedastic-consistent robust estimator in the selected model such as the classical White's sandwich estimator or one of its improved versions (see, e.g. MacKinnon and White, 1985; Long and Ervin, 2000). In Step 3, we generate $\boldsymbol{y}_{n,b}$ from $f_{n,M_{\hat{\jmath}}}(\boldsymbol{y}|X, \vartheta_b, \eta^o_{M_{\hat{\jmath}}})$ for $b = 1, \ldots, B$. This might lead to conservative inference as observed from the simulation study, where the resulting function is denoted as $\tilde{C}_{n,|\hat{\jmath}}(\cdot, \cdot)$.

6 Simulation study

To obtain the post-selection conditional confidence distribution we apply the numerical procedure provided in Section 5 with $B = 1000$. We refer to our proposed method as Post-cc1 when we assume homoscedasticity for the selected model and Post-cc2 when we allow for possible heteroscedasticity in the true generating density and hence increase the variability of the simulated conditional distribution of the sufficient statistics for the interest parameter as explained in Section 5. To obtain the post-selection confidence curves we use $cc_{n|\hat{\jmath}}(\theta) = |1 - 2C_{n|\hat{\jmath}}(\theta, \boldsymbol{Y}_n)|$.

A comparison is made with some of the available methods for producing valid post-selection inference. We compare our post-cc methods to a simultaneous inference method, namely PoSI (Berk et al., 2013) for the homoscedastic case and PoSI (Bachoc et al., 2020) for the heteroscedastic case and a selective inference method with quadratic constraints (Rügamer and Greven, 2018). All these methods have as a target the pseudo-true parameter value in the selected model. The procedure of Rügamer and Greven (2018), see also Tibshirani et al. (2016) and Lee et al. (2016), inverts the cumulative distribution function of a truncated normal distribution in order to obtain the confidence interval bounds. To obtain approximate confidence curves using these other methods, we use the endpoints of a set of $1 - \alpha$ confidence intervals obtained by those methods for a grid of 99 values of $\alpha \in (0, 1)$, spaced by 0.01 using an early concept suggested by Cox (1958). A naive approach that ignores the effect of selection and acts as if the selected model is correct, is presented in the comparison too.

6.1 Homoscedastic normal regression

We simulate data from a linear regression model $Y_i = \sum_{j=1}^{10} \beta_j x_{i,j} + \varepsilon_i, \quad i = 1, \ldots, n = 100$ where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. The true values for the parameters are $\beta = (1.8, -0.3, 1.4, -2.5, \mathbf{0}_6)^\top$ and $\sigma^2 = 1$. The vector $\mathbf{0}_6$ has length 6 and all components are equal to zero. The covariates are generated as $x_{i,1} = 1$ and $(x_{i,2}, \ldots, x_{i,10})^\top \sim N(\mathbf{0}_9, \Omega)$ where Ω is the variance-covariance matrix with correlation everywhere equal to 0.25. We generated 1000 data sets for which a certain correct, but overparametrized, model with parameters $(\beta_1, \ldots, \beta_4, \beta_6)^\top$ is selected. For model selection, we use a backward elimination procedure based on a sequence of t-tests that we call significance hunting at a 5% level with no correction for multiple testing as described in Section 3. This imitates common practice. For all methods except for Post-cc2 the variance σ^2 is treated as known. This allows for a fair comparison with the other methods where this is required for best performance. Therefore Post-cc1 in this simulation setting approximates the confidence distribution in (2). The approximation is due to the Monte-Carlo simulation which is used in the algorithm whenever the exact limits of the truncation are not easily derived. Estimation of the variance and allowing the ε_i, $i = 1, \ldots, n$ to be possibly heteroscedastic is included when using Post-cc2. In this case we obtain conservative confidence curves when we use as $\widetilde{\sigma}(\theta)$ in the modified sampling procedure of Section 5, the estimated standard error in the selected model for $\hat{\beta}_{M_{\hat{\jmath}},r}, \ r = 1, \ldots, 6$.

In this scenario, the targets of inference which are the pseudo-true values coincide with the true parameter values. We show the results for the truly nonzero big effect β_4 and for the truly zero β_6.

For each of the 1000 data sets we obtain confidence curves for the two parameters of interest β_4 and β_6. Figure 2 shows the average confidence curves, the q-q plots for the quantiles of the simulated distribution of $C_n(-2.5, \boldsymbol{Y}_n)$ for β_4 and $C_n(0, \boldsymbol{Y}_n)$ for β_6 versus the expected quantiles of a uniform distribution $\mathcal{U}[0,1]$ and simulated mean coverage probabilities of the $1-\alpha$ confidence intervals with $\alpha \in [0,1]$.

We observe that the average confidence curve's width using our proposed methods are between that of the naive approach which ignores the selection and pretends the selected model to be given beforehand and true, and the width of the other methods. Only our proposed method Post-cc1 and the selective approach lead to confidence distributions that in this setting satisfy condition (R2). Post-cc2 is slightly conservative as expected, that is the price to pay for the uncertainty in the correctness of the model. PoSI (Berk et al., 2013) is shown to be even more conservative, however it seems to fail for the inference of the truly zero $\beta_{M_{\hat{\jmath}},6}$.

In this case, allowing for possible heteroscedasticity in Post-cc2, leads to a loss of power and correspondingly wider confidence intervals. On the contrary, Post-cc1 produces valid post-selection confidence distributions for a correctly specified selected model, as it is an approximated uniformly most powerful confidence distribution.

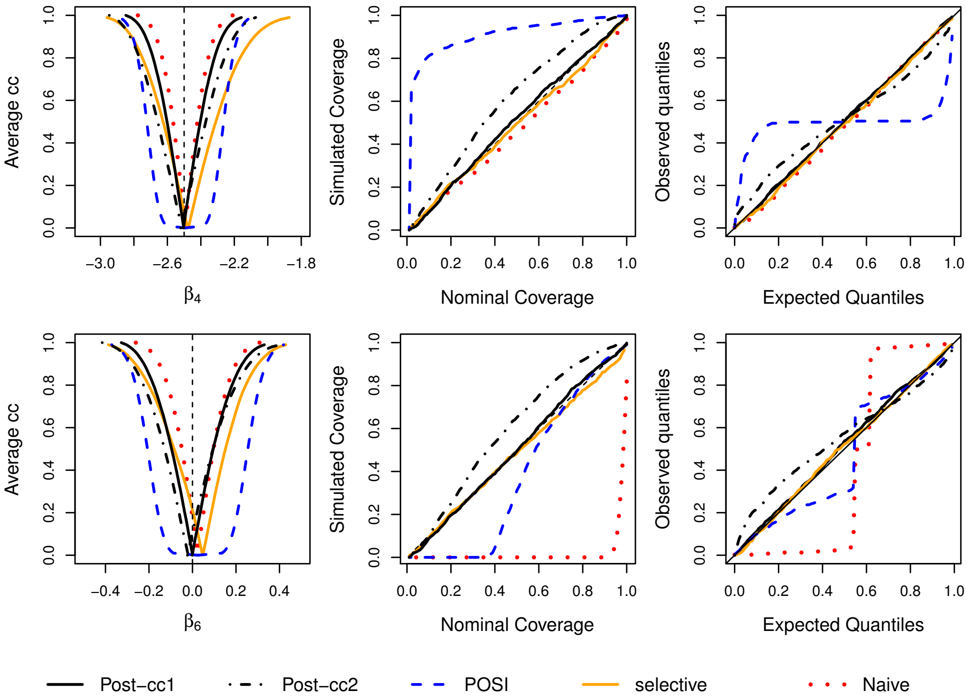

Figure 2: Left: Average confidence curves over 1000 replications for the linear regression parameters β_4 and β_6 when the selection is done by backward elimination based on a sequence of t-tests and the selected linear model is correctly specified. The true parameter values, which in this case coincide with the pseudo-true values, are indicated with a dashed vertical line. Center: Simulated mean coverage of the $1-\alpha$ confidence intervals with $\alpha = [0,1]$, for $\beta_{M_{\hat{\jmath}},4} = -2.5$ and $\beta_{M_{\hat{\jmath}},6} = 0$. Right: Quantiles of the simulated distribution of $C_{n,|\hat{\jmath}}(-2.5, \boldsymbol{Y}_n)$ and $C_{n,|\hat{\jmath}}(0, \boldsymbol{Y}_n)$, for $\beta_{M_{\hat{\jmath}},4}$ and $\beta_{M_{\hat{\jmath}},6}$, respectively, versus expected quantiles of a $\mathcal{U}[0,1]$. Both Post-cc1 and the selective inference method have the correct coverage, though Post-cc1 has narrower intervals and is correctly centered. Post-cc2 is slightly conservative, which is the price to pay for allowing heteroscedasticity.

We remark that the selective method by Rügamer and Greven (2018) produced infinite intervals in, respectively, 7 and 12 data sets for β_4 and β_6, respectively. The problem appears when the estimated parameter is too close to the truncation limits and might indicate that there is little information in the data. This issue was studied for the polyhedral constraints by Kivaranovic and Leeb (2021) who showed that the expected length of the confidence intervals in the worst case scenario is infinite in this type of computation. We removed the data sets with infinite intervals to calculate the selective method's results shown in Figure 2. For β_4 the average confidence curve using the selective inference method is wider than PoSI curves for $1-\alpha > 0.6$, this is caused by data sets whose estimated parameters are close to the truncation limits and produce large but finite confidence intervals (some are even constant over all $1-\alpha$).

We use the same simulation setting to study the coverage of confidence intervals post-selection for a linear combination of the parameters. We consider $x_0^\top\beta$ with $x_0 = (1, 1.5, 0.5, 2, 1.5, 1, 1, 1, 1)^\top$, where the first element corresponds to the intercept. We generate 1000 datasets with the same model specifications as specified above, except that for each dataset the selection procedure in 6.1 is performed and a model is selected. The target of inference is calculated based on the selected model. To obtain the post-selection confidence distributions, the algorithm in section 5 is used with the corresponding sufficient statistics as in section 4.2. Figure 3 shows the simulated versus nominal mean coverage for the pseudo-true values of the linear combination. It is compared to naive inference using a confidence distribution based on the t-statistic for the linear combination in each selected model. Both methods give satisfying results with a slight undercoverage for the naive methods at the most commonly used nominal coverage values. As discussed in Section 2.2, failure of naive methods depend on the value of the parameters to be estimated. In this scenario as the target is a linear combination of the β parameters, it involves several relatively big true values and it is expected for the naive method to perform well. However, as in practice we do not know the true or pseudo-true value of the parameters, it is always preferred to use a method which provides valid inference uniformly over the parameter space.

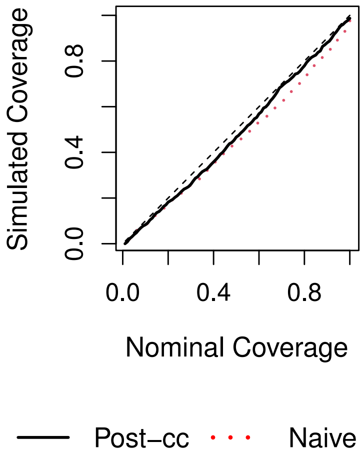

Figure 3: Simulation results with significance hunting at the 5% level. Mean simulated coverage of confidence intervals for a linear combination $x_0^\top\beta$ with covariate vector $x_0 = (1, 1.5, 0.5, 2, 1.5, 1, 1, 1, 1)^\top$. In each replication the target of interest is calculated based on the selected model.

Table 3 in the Appendix provides computation times. Improvements to the algorithm to make it faster are currently under investigation.

6.2 Heteroscedastic normal regression

Data are simulated repeating the setting above but with $\mathrm{Var}(\varepsilon_i) = \sigma_i^2$ non-constant and dependent on one of the covariates of interest. In particular, $\sigma_i^2 = |x_{i,4}^{(3/2)}|$. The sequence of models included in $\mathcal{M}$ is given for the same backward elimination procedure as in the previous setting based on a sequence of t-tests at a 5% level but this time the data are generated to select an underparametrized model in the mean with parameters $(\beta_1, \beta_3, \beta_4, \beta_6)$. To generate the conservative confidence curves using Post-cc2 we use the HC3 estimator (Long and Ervin, 2000) in the selected model. We compare our results with PoSI for heteroscedastic data (Bachoc et al., 2020) and selective inference (Rügamer and Greven, 2018). We remark that the latter method assumes homoscedasticity, therefore our purpose is to assess how it behaves when the assumptions are not correct. We also include the naive approach with t-statistics using the estimated standard error for the coefficient of interest in the selected model ignoring heteroscedasticity (referred as naive) and with the HC3 covariance matrix estimator (referred as naive HC3).

Figure 4 provides the graphical summary for all methods. The average confidence curves using our proposed method are wider than the naive HC3 but much narrower than PoSI. The selective inference method fails in coverage and uniformity of the confidence curves at the pseudo true values. The same conclusion holds for the naive approach. PoSI produces extremely wide confidence curves with coverage close to 1 for all $1-\alpha$ confidence intervals. An interesting analysis is of the inference for $\beta^*_{M_{\hat{\jmath}},6}$. The naive approach leads to a biased point estimator and that also affects the PoSI curves. As our method and that of selective inference by construction are not centered at the maximum likelihood estimator, they are shifted to the averaged pseudo-true value of $\beta^*_{M_{\hat{\jmath}},6}$ (averaged over the 1000 replications). For our method, using the HC3 estimated covariance matrix leads here to over-coverage and the quantiles of the post-selection conservative confidence curves are smaller than the quantiles of $\mathcal{U}[0,1]$ as expected.

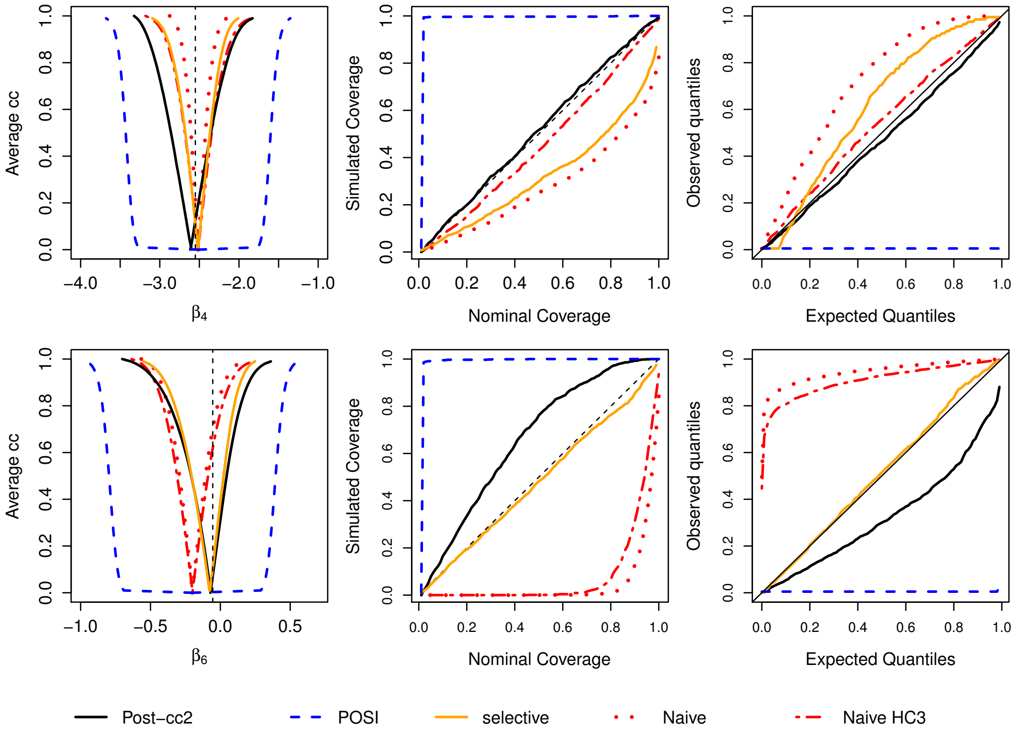

Figure 4: Left: Average confidence curves over 1000 replications for the regression paramaters β_4 and β_6. The data are generated using a linear model with heteroscedastic errors correlated with X_4 and the selected model assumes constant error variance and sets β_2 to zero. The selection procedure is backward elimination based on a sequence of t-tests. The averaged pseudo-true values are indicated with a dashed vertical line. Center: Simulated mean coverage of the $1-\alpha$ confidence intervals with $\alpha = [0,1]$, for $\beta^*_{M_{\hat{\jmath}},4}$ and $\beta^*_{M_{\hat{\jmath}},6}$. Right: Quantiles of the simulated distribution of $\tilde{C}_{n,|\hat{\jmath}}(\beta^*_{M_{\hat{\jmath}},j})$, for $j = 4, 6$ versus quantiles of a $\mathcal{U}[0,1]$. POSI is overly conservative for both cases, while the naive methods have drastic undercoverage for β_6. The selective inference method slightly undercovers. Post-cc2 has correct coverage and narrow intervals for β_4 and is slightly conservative for β_6, still with narrow intervals.

A comparison with sample size=30 in both settings of simulation 1 is provided in the Appendix, see Figures 6 and 7. As expected a smaller sample size leads to wider confidence curves. These results show that the methods also perform well regarding coverage for the smaller sample size of 30.

7 Application: The container terminals data

Lu and Park (2010) studied some factors affecting the productivity of Chinese major container ter-

$$\begin{array}{ll}
M_1: & E(Y|X_{M_1}) = \beta_1\text{yard} + \beta_2\text{Q/C} + \beta_3\text{T/C} + \beta_4\text{Y/T} + \beta_5\text{Length} + \beta_6\text{Depth} \\
M_2: & E(Y|X_{M_2}) = \beta_1\text{yard} + \beta_2\text{Q/C} + \beta_3\text{T/C} + \beta_5\text{Length} + \beta_6\text{Depth} \\
M_3: & E(Y|X_{M_3}) = \beta_2\text{Q/C} + \beta_3\text{T/C} + \beta_5\text{Length} + \beta_6\text{Depth} \\
M_4: & E(Y|X_{M_4}) = \beta_2\text{Q/C} + \beta_3\text{T/C} + \beta_5\text{Length}
\end{array}$$

Table 1: List of structures for the mean in normal linear models in $\mathcal{M}$ obtained by a backward model selection via t-tests at a 5% level in the container terminals data.

minals by linear regression models. The authors performed a backward model selection via t-tests at a 5% level as exemplified in Section 3 and reported the naive p-values for the coefficients in the selected model. We construct conditional post-selection confidence distributions and provide new post-selection estimates, p-values and confidence curves for this study.

The data set consists of 15 observations. The dependent variable is a productivity indicator defined as the annual throughput divided by number of berth. There are 6 covariates, Yard: the yard area in km^2 per berth, Q/C: number of quay cranes per berth, T/C: number of yard cranes per berth, Y/T: number of quay tractors per berth, Length: length of berth in meters , Depth: water depth in meters. Table 1 gives the list of models $M \in \mathcal{M}$, starting with the full standardized model (without intercept) at step 1. In each step the coefficient β_d is discarded, with $d = \arg\min_r\{|\mathcal{T}_{M_s,r}|\ I(|\mathcal{T}_{M_s,r}| \leq t_{\alpha/2})\}$, $\mathcal{T}_{M_s,r} = \hat{\Sigma}^{-1/2}_{M_s,r}\ \hat{\beta}_{M_s,r}$ the t-statistic in model M_s and $t_{\alpha/2}$ equal to the 0.975-quantile of a Student's t-distribution with $n-(p-s+2)$ degrees of freedom. At step 4 all remaining covariates have a corresponding t-statistic larger than the 0.975-quantile of a Student's t-distribution with 11 degrees of freedom so the procedure stops and M_4 is selected.

We obtain the conditional post-selection confidence distributions by method Post-cc1, see Section 5 as there is no evidence of possible heteroscedasticity in the residual plots. Table 2 provides the naive versus the new post-selection estimates and p-values for the coefficients in the selected model M_4. The post-selection estimates are defined as $\tilde{\beta}_{r,n,0.5} = C^{-1}_{n,|\hat{\jmath}}(0.5)$, the median of the conditional post-selection confidence distributions. The p-values for testing $H_0 : \beta_r = 0$ against $H_1 : \beta_r \neq 0$ are obtained as $2\min(C_{n,|\hat{\jmath}}(0, \boldsymbol{y}_n), 1 - C_{n,|\hat{\jmath}}(0, \boldsymbol{y}_n))$. Figure 5 illustrates the confidence curves in comparison with PoSI Berk et al. (2013) and the naive t-distribution. Our method coincides with selective inference of Rügamer and Greven (2018) as the estimates are far from the limits of truncation and both reach optimal inference.

After accounting for the model selection step, the only significant variable at 5% level is the number of yard cranes per berth, T/C. Its post-selection estimate is even bigger than the naive estimate and therefore further away from zero. On the other hand, the whole post-selection confidence curve for T/C is shifted towards zero, while for the covariate Length only the upper part of the post-selection confidence curve includes zero.

Covariate	Standardized Coefficient		p-value	
	Naive	Post-cc1	Naive	Post-cc1
Q/C	0.458	0.360	0.009	0.134
T/C	0.591	0.640	0.002	0.004
Length	-0.350	-0.330	0.004	0.066

Table 2: Naive and post-selection (Post-cc1) estimates and p-values for the coefficients in the selected model M_4 for the container terminals data.

8 Discussion

Our results show that for the normal linear models we can obtain an exact post-selection confidence distribution for the parameter of interest $\theta_{M_{\hat{\jmath}}}$ in the selected model $M_{\hat{\jmath}}$. This approach provides uniformly most powerful hypothesis tests and the tightest valid confidence curves.

As different statistics and methods lead to different confidence curves for the parameter of interest,

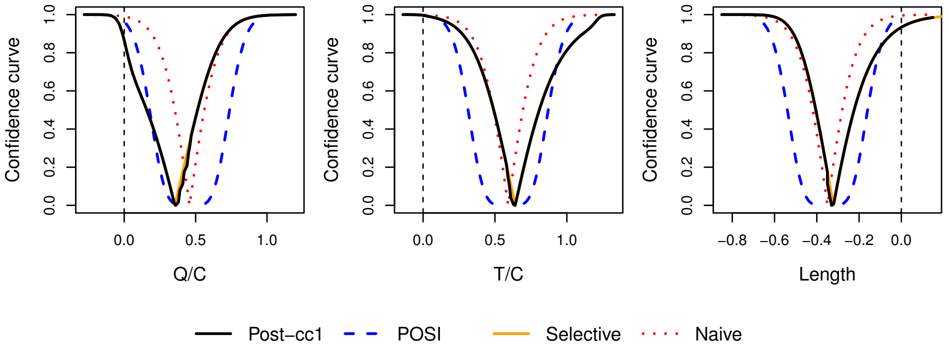

Figure 5: Confidence curves for the standardized coefficients in the selected model M_4 for the container terminals data.

we could formally compare different methods by a measure of tightness given by the area within the confidence curve $\int_\Theta(1-|1-2C_{n,|\hat{\jmath}}(\theta_{M_{\hat{\jmath}}},\boldsymbol{y}_n)|)d\theta_{M_{\hat{\jmath}}}$. In addition to such a tightness number a validity check, which implies proper coverage, is needed as the tightness by itself is not sufficient for comparison. Because confidence distributions and curves might not be easy to obtain explicitly, except for simple models and selection methods, we provide a simulation algorithm too. We show through simulations that our method produces tighter confidence curves than some other valid methods even in the misspecified setting.

In this paper we took the conditional point of view, following the methodology of selective inference, by explicitly conditioning on the event $\boldsymbol{Y}_n \in A_{\hat{\jmath}}$, leading to valid conditional inference using the selected model $M_{\hat{\jmath}}$. Alternatively, by using the law of total probability, for any event B we can write the marginal probability $P(B)=\sum_{M_j\in\mathcal{M}}P\big(B|\boldsymbol{Y}_n\in A_j\big)P\big(\boldsymbol{Y}_n\in A_j\big)$, where the selection region A_j corresponds to selecting model M_j. For confidence intervals and curves, if all of the conditional probabilities provide valid and exact coverage, the same property holds for the marginal statement. In addition, when all conditional statements provide conservative results, again the same result holds for the marginal statement. However, in order to guarantee conservative probabilities, it is not required in a marginal setup that all conditional results are conservative.

Acknowledgements

The authors are grateful to the editor, associate editor and referees for their constructive comments which helped to improve the paper. Funding: Research Foundation Flanders, KU Leuven Research Fund C1-project C16/20/002. The resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. and Csáki, F., editors, *Second International Symposium on Information Theory*, pages 267–281. Akadémiai Kiadó, Budapest.

Bachoc, F., Preinerstorfer, D., and Steinberger, L. (2020). Uniformly valid confidence intervals post-model-selection. *The Annals of Statistics*, 48:440–463.

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

Charkhi, A. and Claeskens, G. (2018). Asymptotic post-selection inference for the Akaike information criterion. *Biometrika*, 105(3):645–664.

Claeskens, G. and Hjort, N. L. (2008). *Model Selection and Model Averaging*. Cambridge University Press, Cambridge.

Cox, D. R. (1958). Some Problems Connected with Statistical Inference. *The Annals of Mathematical Statistics*, 29(2):357–372.

Danilov, D. and Magnus, J. R. (2004). On the harm that ignoring pretesting can cause. *Journal of Econometrics*, 122(1):27–46.

Fithian, W., Sun, D., and Taylor, J. (2017). Optimal inference after model selection. arXiv: 1410.2597.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference and prediction*. Springer, 2nd edition.

Hjort, N. L. and Claeskens, G. (2003). Frequentist model average estimators [with discussion and a rejoinder]. *Journal of the American Statistical Association*, 98:879–899.

Hong, L., Kuffner, T. A., and Martin, R. (2018). On overfitting and post-selection uncertainty assessments. *Biometrika*, 105(1):221–224.

Kabaila, P. (2009). The coverage properties of confidence regions after model selection. *International Statistical Review*, 77(3):405–414.

Kabaila, P., Welsh, A. H., and Abeysekera, W. (2016). Model-averaged confidence intervals. *Scandinavian Journal of Statistics*, 43(1):35–48.

Kivaranovic, D. and Leeb, H. (2021). On the length of post-model-selection confidence intervals conditional on polyhedral constraints. *Journal of the American Statistical Association*, 116:845–857. arXiv: 1803.01665.

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

Lee, J. and Taylor, J. E. (2014). Exact Post Model Selection Inference for Marginal Screening. *arXiv:1402.5596*.

Lehmann, E. L. and Romano, J. P. (2006). *Testing Statistical Hypotheses*. Springer Science & Business Media.

Lehmann, E. L. and Scheffé, H. (1955). Completeness, Similar Regions, and Unbiased Estimation: Part II. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 15(3):219–236.

Lindqvist, B. H. and Taraldsen, G. (2005). Monte Carlo conditioning on a sufficient statistic. *Biometrika*, 92(2):451–464.

Loftus, J. R. (2015). Selective inference after cross-validation. arXiv: 1511.08866.

Loftus, J. R. and Taylor, J. E. (2015). Selective inference in regression models with groups of variables. arXiv: 1511.01478.

Long, J. S. and Ervin, L. H. (2000). Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model. *The American Statistician*, 54(3):217–224.

Lu, B. and Park, N.-K. (2010). A Study on Productivity Factors of Chinese Container Terminals. *Journal of Navigation and Port Research*, 34(7):559–566. Publisher: Korean Institute of Navigation and Port Research.

MacKinnon, J. G. and White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29(3):305–325.

Pya, N. and Wood, S. N. (2015). Shape constrained additive models. *Statistics and Computing*, 25(3):543–559.

Rügamer, D. and Greven, S. (2018). Selective inference after likelihood- or test-based model selection in linear models. *Statistics and Probability Letters*, 140:7–12.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464.

Schweder, T. and Hjort, N. L. (2002). Confidence and Likelihood. *Scandinavian Journal of Statistics*, 29(2):309–332.

Schweder, T. and Hjort, N. L. (2016). *Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

Singh, K., Xie, M., and Strawderman, W. E. (2005). Combining Information from Independent Sources through Confidence Distributions. *The Annals of Statistics*, 33(1):159–183.

Singh, K., Xie, M., and Strawderman, W. E. (2007). Confidence Distribution (CD): Distribution Estimator of a Parameter. *Lecture Notes-Monograph Series*, 54:132–150.

Taylor, J. and Tibshirani, R. (2018). Post-selection inference for ℓ_1-penalized likelihood models.

Canadian Journal of Statistics, 46(1):41–61.

Tian, X., Loftus, J. R., and Taylor, J. E. (2018). Selective inference with unknown variance via the square-root lasso. *Biometrika*, 105(4):755–768.

Tian, X. and Taylor, J. (2017). Asymptotics of Selective Inference. *Scandinavian Journal of Statistics*, 44(2):480–499.

Tian, X. and Taylor, J. (2018). Selective inference with a randomized response. *Annals of Statistics*, 46(2):679–710.

Tibshirani, R. J., Rinaldo, A., Tibshirani, R., and Wasserman, L. (2018). Uniform asymptotic inference and the bootstrap after model selection. *The Annals of Statistics*, 46(3):1255–1287.

Tibshirani, R. J., Taylor, J., Lockhart, R., and Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514):600–620.

White, H. (1994). *Estimation, Inference and Specification Analysis.* Cambridge University Press, Cambridge.

Xie, M.-g. and Singh, K. (2013). Confidence Distribution, the Frequentist Distribution Estimator of a Parameter: A Review. *International Statistical Review*, 81(1):3–39.

Young, G. A. and Smith, R. L. (2005). *Essentials of Statistical Inference.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

A Appendix

Proof of Proposition 1

We condition everywhere on X. Without loss of generality, the combined design matrix is reorganized as $X = (X_{M_{\hat{\jmath}}}, X^c_{M_{\hat{\jmath}}})$. Under the working selected model, $\beta^*_{M_{\hat{\jmath}}} = (X^\top_{M_{\hat{\jmath}}} X_{M_{\hat{\jmath}}})^{-1} X^\top_{M_{\hat{\jmath}}} \mu$ and we take $\theta^* = \beta^*_{M_{\hat{\jmath}},1}/\sigma^2$. Let $T = X^\top_{M_{\hat{\jmath}},1} \boldsymbol{Y}_n$, where $X_{M_{\hat{\jmath}},1}$ is the first column of X and $U_{\mathcal{M}} = X^\top_{-1} \boldsymbol{Y}_n$ where X_{-1} is the design matrix without the first column. Note that when σ^2 is known or when we plug in an independent estimator, $U_{\mathcal{M}}$ does not include $\boldsymbol{Y}_n^\top \boldsymbol{Y}_n$. With $U_{\hat{\jmath}} = U(\boldsymbol{Y}_n; X_{M_{\hat{\jmath}}})$ the subvector of $U_{\mathcal{M}}$ consisting of the sufficient statistics for the nuisance parameters $\eta^*_{\hat{\jmath}}$ in the selected model and with observed value $u_{\hat{\jmath}}$, we can reorganize $U_{\mathcal{M}} = \big(U_{\hat{\jmath}}, U^c_{\hat{\jmath}}\big)^\top$. For the nuisance parameters not appearing in the selected model $\eta^{c*}_{\hat{\jmath}} = 0$, a vector of all zeros.

In addition, denote $P_{X_{M_{\hat{\jmath}},-1}} = X_{M_{\hat{\jmath}},-1}(X^\top_{M_{\hat{\jmath}},-1} X_{M_{\hat{\jmath}},-1})^{-1} X^\top_{M_{\hat{\jmath}},-1}$ a projection matrix, $A = X^\top_{M_{\hat{\jmath}},1} X_{M_{\hat{\jmath}},1} - X^\top_{M_{\hat{\jmath}},1} P_{X_{M_{\hat{\jmath}},-1}} X_{M_{\hat{\jmath}},1}$ and $C = X^\top_{M_{\hat{\jmath}},1} X_{M_{\hat{\jmath}},-1}(X^\top_{M_{\hat{\jmath}},-1} X_{M_{\hat{\jmath}},-1})^{-1}$. Then, the parameter of interest is $\theta^* = 1/\sigma^2 A^{-1}[X^\top_{M_{\hat{\jmath}},1}\mu - C X^\top_{M_{\hat{\jmath}},-1}\mu]$.

The joint distribution of $(T, U_{\hat{\jmath}}, U^c_{\hat{\jmath}})$ is normal with mean $X^\top \mu$ and variance $\sigma^2(X^\top X)$. The conditional distribution of $T | U_{\hat{\jmath}} = u_{\hat{\jmath}}$ is again a normal distribution of the form

$$f_{T|U_{\hat{\jmath}}}(t; X_{M_{\hat{\jmath}}}, \theta) = \exp\big\{ -\frac{1}{2\sigma^2} t^2 + \theta t + \frac{1}{\sigma^2} A^{-1} C^\top u_{\hat{\jmath}} t - \kappa'(\theta, X_{M_{\hat{\jmath}}}) \big\}$$

with $\kappa'(\cdot)$ a generic function of the exponential family. When σ^2 is known the density can be rewritten as $f_{T|U_{\hat{\jmath}}}(t; X_{M_{\hat{\jmath}}}, \theta) = h'(t) \exp\big\{\theta t - \kappa'(\theta, X_{M_{\hat{\jmath}}})\big\}$. If an estimator $\hat{\sigma}^2$ is plugged in which is estimated independently of T and $U_{\mathcal{M}}$, the conditional distribution of $T|(U_{\hat{\jmath}}, \hat{\sigma}^2)$ equals the distribution of $T|U_{\hat{\jmath}}$.

The fact that the conditional distribution of $T|U_{\hat{\jmath}} = u_{\hat{\jmath}}$ constitutes a one-parameter exponential family and that $U_{\hat{\jmath}}$ is a complete sufficient statistic for $\eta^*_{\hat{\jmath}}$ parallels the framework for obtaining the Neyman–Pearson optimality of conditional tests for θ^*, the parameter of interest, when $U_{\hat{\jmath}} = u_{\hat{\jmath}}$ is given, see Theorem 4.4.1 of Lehmann and Romano (2006).

Conditioning additionally on $U^c_{\hat{\jmath}}$ keeps the same generic form of the working conditional density, and with some algebra we see that the functions $h'(\cdot)$ and $\kappa'(\cdot)$ now depend on the combined design matrix X. So the density of $T \mid U_{\mathcal{M}} = u_{\text{obs}}$ is still depending on the single parameter θ^* and is denoted by

$$f_{T|U_{\mathcal{M}}}(t; X, \theta) = h'(t, X) \exp\big\{\theta t - \kappa'(\theta, X)\big\}. \tag{5}$$

After selection, without loss of generality, assume that the domain is of the following form $\text{dom}(T|U_{\mathcal{M}} = u_{\text{obs}}, \boldsymbol{Y}_n \in A_{\hat{\jmath}}) = \{t = t(y; X_{M_{\hat{\jmath}}}) \in \mathbb{R} : a \leq t \leq b\}$, with a and b determined by the specificities of the selection procedure and fixed after conditioning on $U_{\mathcal{M}} = u_{\text{obs}}$. Then $T|(U_{\mathcal{M}} = u_{\text{obs}}, \boldsymbol{Y}_n \in A_{\hat{\jmath}})$ follows a truncated exponential family distribution of the form

$$f_{T|U_{\mathcal{M}},\hat{\jmath}}(t; X, \theta) = \frac{h'(t, X) \exp\big\{\theta t - \kappa'(\theta, X)\big\} I(a \leq t \leq b)}{G_{T|U_{\mathcal{M}}}(b; \theta) - G_{T|U_{\mathcal{M}}}(a; \theta)}, \tag{6}$$

where $G_{T|U_{\mathcal{M}}}$ is the cumulative distribution function of $T|U_{\mathcal{M}}$ and $I(\cdot)$ is the indicator function. For any $\theta_2 > \theta_1 \in \Theta$, the likelihood ratio $f_{T|U_{\mathcal{M}},\hat{\jmath}}(t; X_{M_{\hat{\jmath}}}, \theta_2)/\ f_{T|U_{\mathcal{M}},\hat{\jmath}}(t; X_{M_{\hat{\jmath}}}, \theta_1)$ is equal to

$$LR(\theta_1, \theta_2, t) = \exp\left\{\frac{\kappa'(\theta_1, X)}{\kappa'(\theta_2, X)}\right\} \cdot \frac{[G_{T|U_{\mathcal{M}}}(b; \theta_1) - G_{T|U_{\mathcal{M}}}(a; \theta_1)]}{[G_{T|U_{\mathcal{M}}}(b; \theta_2) - G_{T|U_{\mathcal{M}}}(a; \theta_2)]} \cdot \exp\big\{(\theta_2 - \theta_1)t\big\}.$$

The first and second factors of the right-side are constant in t while the third factor is increasing in t for every $\theta_2 > \theta_1 \in \Theta$. Therefore $LR(\theta_1, \theta_2, t)$ is everywhere increasing.

Finally, denote the cumulative distribution function of (6) by $G_{T|U_{\mathcal{M}},\hat{\jmath},\theta}$. For any loss function $\text{loss}(\theta, C) = \int B(\theta' - \theta)\ dC(\theta', Y)$, with B nondecreasing on the positive half-axis, nonincreasing on the negative half-axis and $B(0) = 0$, by Schweder and Hjort (2016, Theorem 5.10), the confidence distribution $C_{n,|\hat{\jmath}}(\theta, \boldsymbol{Y}_n) = 1 - G_{T|U_{\mathcal{M}},\hat{\jmath},\theta}$, which is based on a sufficient statistic in which the likelihood ratio is everywhere increasing, minimizes the confidence loss uniformly over Θ. The same applies to any fixed limits that truncate the domain of $T|(U_{\mathcal{M}} = u_{\text{obs}}, \boldsymbol{Y}_n \in A_{\hat{\jmath}})$. □

Proof of Proposition 2

We condition everywhere on X. As in proof of Proposition 1, we reorganize $U_{\mathcal{M}} = \big(U_{\hat{\jmath}}, U_{\hat{\jmath}}^c\big)^\top$, with $U_{\hat{\jmath}} = U(\boldsymbol{Y}_n; X_{M_{\hat{\jmath}}})$ the subvector of $U_{\mathcal{M}}$ consisting of the sufficient statistics for the nuisance parameters in the selected model $\eta_{\hat{\jmath}}^*$ and with observed value $u_{\hat{\jmath}}$. Note that $\eta_{\hat{\jmath}}^{c*} = 0$. Denote $P_{X_{M_{\hat{\jmath}}}} = X_{M_{\hat{\jmath}}}(X_{M_{\hat{\jmath}}}^\top X_{M_{\hat{\jmath}}})^{-1} X_{M_{\hat{\jmath}}}^\top$ the projection matrix using $X_{M_{\hat{\jmath}}}$.

We reparametrize the joint density of $\boldsymbol{Y}_n$ given X as

$$\exp\left\{-\frac{1}{2\sigma^2}\boldsymbol{y}_n^\top \boldsymbol{y}_n + \frac{1}{\sigma^2}\boldsymbol{y}_n^\top P_{X_{M_{\hat{\jmath}}}}\mu + \frac{1}{\sigma^2}\boldsymbol{y}_n^\top (I - P_{X_{M_{\hat{\jmath}}}})\mu - \frac{1}{2\sigma^2}\mu^\top P_{X_{M_{\hat{\jmath}}}}(I - P_{X_{M_{\hat{\jmath}}}})\mu + \frac{n}{2}\log(2\pi\sigma^2)\right\},$$

which can be rewritten as a exponential family in its natural parametrization:

$$\exp\left\{-\frac{1}{2\sigma^2}\boldsymbol{y}_n^\top \boldsymbol{y}_n + \frac{1}{\sigma^2}(X_{M_{\hat{\jmath}}}^\top \boldsymbol{y}_n)^\top \beta_{M_{\hat{\jmath}}} + \frac{1}{\sigma^2}\boldsymbol{y}_n^\top (I - P_{X_{M_{\hat{\jmath}}}})\mu - \frac{1}{2\sigma^2}X_{M_{\hat{\jmath}}}^\top \beta_{M_{\hat{\jmath}}}^\top (I - P_{X_{M_{\hat{\jmath}}}})\mu + \frac{n}{2}\log(2\pi\sigma^2)\right\}, \tag{7}$$

with natural parameters $(\beta_{M_{\hat{\jmath}}}^\top/\sigma^2, -1/(2\sigma^2), \mu^\top/\sigma^2)$ and sufficient statistics
$(\widetilde{T}(\boldsymbol{Y}_n; X_{M_{\hat{\jmath}}}), V) = (X_{M_{\hat{\jmath}}}^\top \boldsymbol{Y}_n, \boldsymbol{Y}_n^\top \boldsymbol{Y}_n, (I - P_{X_{M_{\hat{\jmath}}}})\boldsymbol{Y}_n)$. Now, reorder $\widetilde{T}(\boldsymbol{Y}_n; X_{M_{\hat{\jmath}}}) = (T, U_{\hat{\jmath}}^\top)^\top$ such that the parameter of interest $\theta = \beta_{M_{\hat{\jmath}},1}^\top/\sigma^2$ or $\theta = -1/(2\sigma^2)$.

Because the parameter space for $(\theta^*, \eta_{\hat{\jmath}}^{*\top}, \mu^\top)^\top$ contains an open rectangle in $\mathbb{R}^{n+p_{M_{\hat{\jmath}}}+1}$ with $p_{M_{\hat{\jmath}}} \leq p$ and the joint density of $\boldsymbol{Y}_n$ under reparametrization in (7) represents an exponential family density in its natural parameterization, it follows that $(U_{\hat{\jmath}}^\top, V^\top)^\top$ is a complete sufficient statistic for $(\eta_{\hat{\jmath}}^{*\top}, \mu^\top)^\top$ (Lehmann and Romano, 2006, Th. 4.3.1). By Lemma 2.7.2 of Lehmann and Romano (2006), T conditioned on $(U_{\hat{\jmath}} = u_{\hat{\jmath}}, V = v_{\text{obs}})$ is again in the exponential family. The density of $T|(U_{\hat{\jmath}} = u_{\hat{\jmath}}, V = v_{\text{obs}})$ is hence denoted by $h'(t(\boldsymbol{y}_n; X_{M_{\hat{\jmath}}}))\exp\big\{\theta t(\boldsymbol{y}_n; X_{M_{\hat{\jmath}}}) - \kappa'(\theta, X_{M_{\hat{\jmath}}})\big\}$ for some functions $h'(\cdot)$ and $\kappa'(\cdot)$. Due to the sufficiency of $U_{\hat{\jmath}}$ and V, the above density depends on θ^* but not on η^* or μ. Conditioning additionally on $U_{\hat{\jmath}}^c$ keeps the same generic form of the working conditional density. So the density of $T \mid (U_{\mathcal{M}} = u_{\text{obs}}, V = v_{\text{obs}})$ is still depending on the single parameter θ^* and is denoted by the right-side of (5). We continue now as in the proof of Proposition 1 by replacing $T \mid (U_{\mathcal{M}} = u_{\text{obs}})$ by $T \mid (U_{\mathcal{M}} = u_{\text{obs}}, V = v_{\text{obs}})$ and $\text{dom}(T|U_{\mathcal{M}} = u_{\text{obs}}, \boldsymbol{y}_n \in A_{\hat{\jmath}})$ by $\text{dom}(T|U_{\mathcal{M}} = u_{\text{obs}}, V = v_{\text{obs}}, \boldsymbol{y}_n \in A_{\hat{\jmath}})$ which is fixed after conditioning on $U_{\mathcal{M}} = u_{\text{obs}}$. □

Sample size effect: Comparison simulation studies

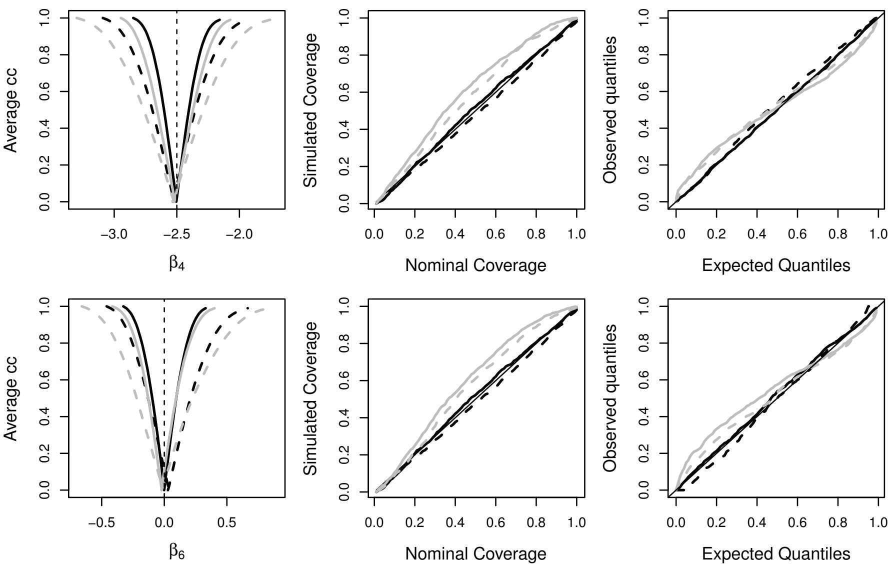

Figure 6: Simulation under homoscedasticity. Effect of sample size. Confidence curves, mean simulated coverage of confidence intervals and quantiles of simulated distributions of confidence distributions at true value vs expected quantiles of a $\mathcal{U}[0,1]$ for the parameters of interest when the selected linear model is correctly specified with sample size=100 (solid line) and 30 (dashed line). The darker colored confidence curves are obtained using the sampling approach for correctly specified models in Section 5 and the lighter colored when we allow for possible misspecification and use the robust HC3 estimator for the covariance matrix. The coverage results are accurate also for $n = 30$, while as expected, a tighter confidence distribution is obtained for larger sample sizes.

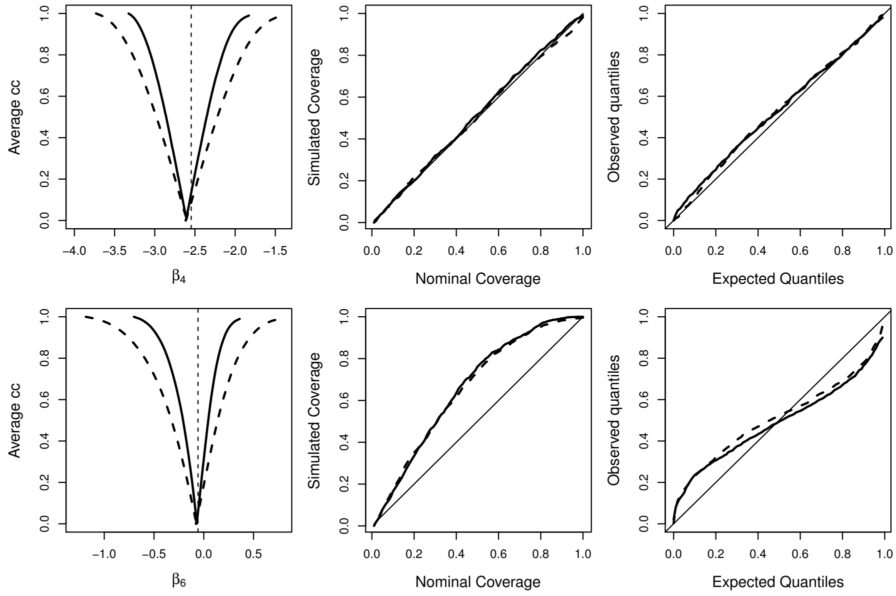

Figure 7: Simulation under heteroscedasticity. Effect of sample size. Confidence curves, mean simulated coverage of confidence intervals and quantiles of simulated distributions of confidence distributions at pseudo-true value versus expected quantiles of a $\mathcal{U}[0,1]$ for the parameters of interest when the selected linear model is misspecified with sample size=100 (solid line) and 30 (dashed line). We adjust the sampling approach with the robust HC3 estimator for the covariance matrix. The coverage results are accurate for β_4 and conservative for β_6, for both sample sizes, while as expected, a tighter confidence distribution is obtained for larger sample sizes.

Computation times

Table 3 shows the average run-time over 5 replications to obtain a confidence curve for a regression parameter of interest, $\beta_{M_{\hat{\jmath}},6}$, after the data were generated with the settings of simulation 6.1. (*Computations were performed in a single core AMD Ryzen 5 5500U 2.10 GHz, parallel computing speeds up computations.)

Post-cc*	7. 711 (sd=2.729) mins
selective	1.792 (sd=0.029) mins
PoSI	0.638 (sd=0.079) secs

Table 3: Average run-time over 5 replications to obtain a confidence curve in simulation 6.1.