Optimal finite sample post-selection confidence distributions in generalized linear models

Andrea C. Garcia-Angulo[a] **and Gerda Claeskens**[b]

[a]Facultad de Ciencias Naturales y Matemáticas, Escuela Superior Politécnica del Litoral, ESPOL; [b]ORStat and Leuven Statistics Research Centre, KU Leuven, Naamsestraat 69, Box 3555, B-3000 Leuven, Belgium.
acgarcia@espol.edu.ec; gerda.claeskens@kuleuven.be

Abstract

Uniformly most powerful confidence distributions are obtained for parameters in selected models of the exponential family. A conditioning on the selection event as well as on the sufficient statistics of nuisance parameters guarantees valid post-selection inference. Optimal confidence intervals are obtained directly from the confidence distribution without requiring an inversion of pivotal quantities. Simulations showcase that the method works also when all models are misspecified.

Keywords: Confidence distribution, confidence interval, exponential family, model selection, post-selection inference, sufficiency

1 Introduction

While variable selection in generalized linear models is now standard, the construction of valid post-selection inference is still not commonplace. Via the concept of confidence distributions (see, e.g., Schweder and Hjort, 2016) which summarize all information about the power of tests, p-values and confidence intervals at all levels, we approach the inference on model parameters after a model has been selected using the same data. In linear models this method leads to finite sample exact and uniformly most powerful results, see Garcia-Angulo and Claeskens (2022). In this paper we extend the methodology to be applicable to the class of generalized linear models. We test the method on simulated data for selecting Poisson and logistic regression models and we re-analyse a published analysis using logistic regression but now obtaining valid inference results after selection.

This method can be framed as selective inference (see Lee et al., 2016, for Gaussian data) where a pivotal quantity is used conditional on a selected model by characterizing the event of selection. For lasso selection this selection event corresponds to certain polyhedral regions while for selection by the Akaike information criterion in likelihood models quadratic regions results (Charkhi and Claeskens, 2018). Rügamer and Greven (2018) consider the construction of selection regions and conditional inference for likelihood, or test-based model selection in a linear model. For non-Gaussian data much fewer results exist for selective post-selection inference. Some asymptotic results are obtained by Tian and Taylor (2017) and a more powerful method by the use of randomization is provided by Tian and Taylor (2018). Taylor and Tibshirani (2018) discuss extensions to generalized regression models by using one-step approximations and asymptotic normality of the resulting estimators. Tibshirani et al. (2018) obtain asymptotic uniform inference for linear regression models with non-Gaussian errors. For other ways of performing valid inference that do not condition on the selection, but consider some orthogonality conditions instead, see for example Belloni et al. (2016) for the case of ℓ_1-regularized estimation in generalized linear models. None of these papers has considered the use of confidence distributions to get full inferential results and no finite sample results were obtained.

2 Models, sufficient statistics and selection regions

Let $\boldsymbol{Y}_n = (Y_1, \ldots, Y_n)^\top$ be a n-dimensional vector of independent random variables. In an exponential family model, each Y_i has a density or probability mass function (pmf) of the form

$$f_i(y_i; \zeta_i, \phi) = \exp\left\{\frac{y_i\zeta_i - b(\zeta_i)}{\phi} + c(y_i, \phi)\right\}, \tag{1}$$

where b and c are known functions, ϕ is a scale parameter and ζ_i is the canonical parameter of the exponential family which might be different for each $i = 1, \ldots, n$. Denote $b'(\zeta_i) = \partial b(\zeta_i)/\partial\zeta_i$ and $b''(\zeta_i) = \partial^2 b(\zeta_i)/\partial\zeta_i^2$. For members of the exponential family distributions with densities or probability mass functions expressed as (1) it can be shown that $E(Y_i) = b'(\zeta_i)$ and $\mathrm{Var}(Y_i) = \phi b''(\zeta_i)$.

For p-vectors of covariates x_i with $i = 1, \ldots, n$, define the $n \times p$ full rank design matrix $X = (x_1^\top, \ldots, x_n^\top)^\top$ and $\beta = (\beta_1, \ldots, \beta_p)^\top$ the corresponding parameter vector. For fixed regressors, the responses are independent though their means will differ. For a random design the vectors $(Y_i, x_i^\top)$, $i = 1, \ldots, n$ are assumed independent and identically distributed and in this case in all expressions a conditioning on X takes place, though this is not always explicitly indicated in the notation.

The class of generalized linear models (GLM) with canonical link function $\tilde{\mathrm{g}} = (b')^{-1}$ specifies the canonical parameter as $\zeta_i = \tilde{\mathrm{g}}(E(Y_i|x_i)) = x_i^\top\beta$ such that the joint density or probability mass function of $\boldsymbol{Y}_n|X$ has the form

$$f_n(\boldsymbol{y}_n|X, \beta, \phi) = \prod_{i=1}^n f_i(y_i|x_i, \zeta_i, \phi) = \prod_{i=1}^n \left[\exp\left\{\frac{y_i x_i^\top\beta - b(x_i^\top\beta)}{\phi} + c(y_i, \phi)\right\}\right]. \tag{2}$$

When ϕ is known, it immediately follows from (2) and the factorization criterion for sufficiency, that $\sum_{i=1}^n x_i Y_i$ is a sufficient statistic for β/ϕ. When ϕ requires estimation, $c(y_i, \phi)$ contains information about the sufficient statistic for ϕ. Table 1 gives a list of the most used GLM with canonical links and their natural parameters and sufficient statistics. Working with the canonical link function allow us to rewrite (2) in what is called the natural parametrization of exponential family distributions,

$$f_n(\boldsymbol{y}_n|X, \pi(\beta, \phi)) = h(\boldsymbol{y}_n)\ \exp\left\{\pi(\beta, \phi)^\top\widetilde{T}(\boldsymbol{y}_n; X) - \kappa(\pi(\beta, \phi))\right\}, \tag{3}$$

with a k-dimensional vector of natural parameters $\pi(\beta, \phi) = (\pi_1(\beta, \phi), \ldots, \pi_k(\beta, \phi))^\top$ and (conditional on X) the corresponding vector of sufficient statistics $\widetilde{T}(\boldsymbol{y}_n; X) = (T_1(\boldsymbol{y}_n; X), \ldots, T_k(\boldsymbol{y}_n; X))^\top$, with $k = p + 1$ when ϕ requires estimation, and with $k = p$ when ϕ is

Distribution	Canonical link $\tilde{\mathrm{g}}(E(Y_i\|x_i)) = x_i^\top\beta$	Natural parameters	Sufficient statistics
Normal: $N(\mu_i, \sigma^2)$	$\mu_i = x_i^\top\beta$	$(-1/2, \beta^\top)^\top/\sigma^2$	$(Y^\top Y, Y^\top X)^\top$
Inverse Gaussian(μ_i, γ)	$-1/(2\mu_i^2) = x_i^\top\beta$	$(-\gamma/2, \gamma\beta^\top)^\top$	$(\sum_{i=1}^n Y_i^{-1}, Y^\top X)^\top$
Exponential(λ_i)	$-\lambda_i = x_i^\top\beta$	β	$X^\top Y$
Gamma(v, λ_i)	$-\lambda_i/v = x_i^\top\beta$	$(v, v\beta^\top)^\top$	$(\sum_{i=1}^n \log(Y_i), Y^\top X)^\top$
Poisson(λ_i)	$\log(\lambda_i) = x_i^\top\beta$	β	$X^\top Y$
Binomial: $\mathrm{Bin}(\mathrm{n_i}, \mathrm{p_i})$	$\log(\frac{n_i p_i}{n_i(1-p_i)}) = x_i^\top\beta$	β	$X^\top Y$

Table 1: Most used generalized linear models with canonical links and their natural parameters and sufficient statistics.

known. A slightly more general case allows to specify $\zeta_i = \tilde{\mathrm{h}}(E(Y_i|x_i))$, where $\tilde{\mathrm{h}}$ is any monotone function that is linear in the parameter β. In this case, the working model also yields the natural parametrization of the exponential family as in (3).

We first consider the case that a single regression coefficient is the focus for inference. The natural parameter vector is split as $\pi(\beta, \phi) = (\theta, \eta^\top)^\top$ and we denote the vector of sufficient statistics $\widetilde{T}(\boldsymbol{Y}_n; X) = (T(\boldsymbol{Y}_n; X), U^\top(\boldsymbol{Y}_n; X))^\top$. The first component $T = T(\boldsymbol{Y}_n; X)$ is the sufficient statistic for θ, while the other components form the vector of sufficient statistics for what we call the nuisance parameter vector η. We use this parametrization to obtain post-selection confidence distributions for θ based on the conditional distribution of T given U.

Model selection takes place in order to create a more parsimonious model with fewer covariates. Using a selection method a best model is chosen from the model set $\mathcal{M} = \{M_1, \ldots, M_m\}$ which consists of a fixed number m models.

Each model selection method imposes a partitioning of $\mathcal{Y}$, the sample space of the data. That is, $\mathcal{Y} = \cup_{j=1}^m A_j$ with $A_k \cap A_l = \emptyset$ if $k \neq l$. Each of these regions A_j is connected to one model $M_j \in \mathcal{M}$ such that selecting model M_j is equivalent with $\boldsymbol{Y}_n \in A_j$. All models in $\mathcal{M}$ are assumed to have a nonzero selection probability implying that each A_j with $j \in \{1, \ldots, m\}$ is nonempty. We make the assumption that the selection regions can be described using the sufficient statistics. The selected model is denoted by $M_{\widehat{j}}$ to stress the randomness in the selection in the notation.

Example: Poisson regression. The logarithm is the canonical link function such that $E(Y_i|x_i) = \exp(x_i^\top\beta)$. The working density in (2) takes a known $\phi = 1$, $b = \exp$ and $c(y_i, \phi) = -\log(y_i!)$. It can be rewritten as in (3) with $h(\boldsymbol{y}_n) = \prod_{i=1}^n \exp\{c(y_i, \phi)\}$, $\kappa(\pi(\beta, \phi)) = \sum_{i=1}^n b(x_i^\top\beta)$ and a p-variate sufficient statistic for $\pi(\beta, \phi) = \beta$ is $\widetilde{T}(\boldsymbol{y}_n; X) = \sum_{i=1}^n x_i Y_i$.

Let $\mathcal{M} = \{M_1, \ldots, M_m\}$ be the model selection set where each model M_j is a Poisson regression model with canonical log-link function and design matrix X_{M_j} with $|M_j|$ columns, which denotes the number of parameters that is to be estimated in model M_j. Models in $\mathcal{M}$ are not necessarily nested. The probability mass function for model M_j, can be expressed in terms of the sufficient statistics as

$$f_n(\boldsymbol{y}_n|X_{M_j}, \beta_{M_j}) = \left(\prod_{i=1}^n \frac{1}{y_i!}\right) \exp\left\{\beta_{M_j}^\top \sum_{i=1}^n x_i y_i - \sum_{i=1}^n \exp(x_i^\top \beta_{M_j})\right\}.$$

Let $M_{\widehat{\jmath}}$ be selected by AIC. Since $AIC(M_j) = -2\log f_n(\boldsymbol{y}_n|X_{M_j}, \widehat{\beta}_{M_j}) + 2|M_j|$, using the model's maximized log-likelihood value, it is readily seen that also $A_{\widehat{\jmath}}$ can be expressed in terms of the sufficient statistics. The selection yields the following selection region for the observed sample: $A_{\widehat{\jmath}} = \{\boldsymbol{y}_n \in \mathbb{R}^n : \text{AIC}(M_{\widehat{\jmath}}) < \text{AIC}(M_k), \text{for all } M_k \in \mathcal{M} \setminus M_{\widehat{\jmath}}\}$.

3 Post-selection conditional confidence distributions

3.1 Uniformly most powerful confidence distributions

Garcia-Angulo and Claeskens (2022) define a conditional post-selection confidence distribution as a confidence distribution (Schweder and Hjort, 2002, 2016; Singh et al., 2005; Xie and Singh, 2013) that is restricted to the area of selection instead of on the whole sample space and there is a conditioning on the selected model. In order to take possible model misspecification into account, pseudo-true parameters (White, 1994) are used instead of true parameter values. Such values are defined to minimize the Kullback-Leibler divergence between the model density and the true data generating density, the latter is denoted by g. Thus for a possibly misspecified model with density function $f_n(\boldsymbol{y}_n|X, \pi(\beta, \phi))$ and parameter of interest renamed to θ and the remaining, nuisance parameters renamed to η, the pseudo-true parameter value in a model M is defined as $(\theta^*_M, \eta^{*\top}_M) = \arg\max_{(\theta,\eta^\top)} E_g[\log f_n(\boldsymbol{Y}_n|X_M, \pi(\beta_M, \phi))]$ where the expectation is with respect to the true but unknown density of $\boldsymbol{Y}_n$ and the model M is considered given.

Conditional on the selected model $M_{\widehat{\jmath}}$ with pseudo-true parameter vector $(\theta^*_{M_{\widehat{\jmath}}}, \eta^{*\top}_{M_{\widehat{\jmath}}})$, Garcia-Angulo and Claeskens (2022) defined a function

$$C_{n|\widehat{\jmath}} : \Theta \times A_{\widehat{\jmath}} \to [0,1] : (\theta, \boldsymbol{Y}_n) \mapsto C_{n|\widehat{\jmath}}(\theta, \boldsymbol{Y}_n)$$

to be a conditional post-selection confidence distribution if

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

(ii) whatever the value of the pseudo-true parameter vector, considered as a function of $\boldsymbol{Y}_n$ taking values in $A_{\widehat{\jmath}}$, $C_{n|\widehat{\jmath}}(\theta^*_{M_{\widehat{\jmath}}}, \boldsymbol{Y}_n) \sim \mathcal{U}[0,1]$, a uniform distribution.

By explicitly using properties of the normal distribution function, Garcia-Angulo and Claeskens (2022) constructed such a conditional post-selection confidence distribution for selection among normal linear models and proved it to be finite sample uniformly most powerful. The novelty of this paper is that we extend such property to selection in an exponential family class of models. We prove below that theoretical finite sample results can be obtained for continuous such distributions under some assumptions. We provide also answers for discrete distributions and for cases where all models are wrong.

Optimality properties of exponential families were extended to confidence distributions by Schweder and Hjort (2016, Ch. 5). A confidence distribution is uniformly optimal (most powerful) if for every loss function B, nondecreasing on the positive half-axis, nonincreasing on the negative half-axis and $B(0) = 0$, $\text{loss}(\theta_a, C_{opt}) \leq \text{loss}(\theta_a, C)$ for any other C, for every value θ_a (Schweder and Hjort, 2016, Def 5.9).

We first consider the scenario where there might be misspecified models but the selected model is minimal true or overparametrized. In this case exact finite sample results hold. An overspecified model M_j contains the minimal true model and it is overparametrized if there are some zero components in the parameter vector β_{M_j}.

This is precisely the same scenario as in Tian and Taylor (2018) and Fithian et al. (2014) for the general exponential family (disregarding the Gaussian case) where selective tests based on the conditional distributions of the sufficient statistics are only valid under the model assumptions. This is not a strong assumption since many model selection methods will overselect, including the lasso approach and efficient selection methods such as the Akaike information criterion, for example.

In this case we can obtain optimal exact finite sample inference for the parameter of interest. We first state the result for continuous distributions. Discrete distributions require a continuity correction, as is explained below. We here extend Proposition 2 of Garcia-Angulo and Claeskens (2022), which proved optimality for selecting normal linear models, to generalized linear models.

For $M_{\widehat{\jmath}}$ the minimal true or overspecified selected model using the data $\boldsymbol{Y}_n$ with a continuous exponential family density, and the design matrix $X_{M_{\widehat{\jmath}}}$, the univariate parameter of interest is θ with parameter space Θ and sufficient statistic $T = T(\boldsymbol{Y}_n, X_{M_{\widehat{\jmath}}})$ with observed value t_{obs}. The $(|M_{\widehat{\jmath}}| - 1)-$dimensional vector of nuisance parameters is η with sufficient statistic $U = U(\boldsymbol{Y}_n, X_{M_{\widehat{\jmath}}})$. Define by $U_{\mathcal{M}} = U_{\mathcal{M}}(\boldsymbol{Y}_n, X)$ the combined vector of sufficient statistics for the nuisance parameters η^* in $\mathcal{M}$, removing duplicates, with observed value u_{obs}. Since the decision on which parameter to perform inference on and which other parameters are nuisance happens after the model selection step, a conditioning on the selected model takes place and the inference is different from classical inference where the model is given beforehand. Section 7 contains the proof.

Proposition 1. *We condition on $M_{\widehat{\jmath}}$ being an overspecified selected model with parameters $(\theta, \eta^\top)$ from a set of continuous exponential family models $\mathcal{M}$, which is equivalently expressed as $\boldsymbol{Y}_n \in A_{\widehat{\jmath}}$, the corresponding selection region which can be expressed in terms of the sufficient statistics $(T(\boldsymbol{Y}_n; X_{M_{\widehat{\jmath}}}), U_{\mathcal{M}}^\top)$ for the model parameters in $\mathcal{M}$. If the parameter space for $(\theta, \eta^\top)$ contains an open rectangle and the sample space region does not depend on the parameters,*

(i) the conditional post-selection confidence distribution:

$$C_{n,[\widehat{\jmath}} : \Theta \times A_{\widehat{\jmath}} \to [0,1] : (\theta, \boldsymbol{Y}_n) \mapsto P\big(T(\boldsymbol{Y}_n; X_{M_{\widehat{\jmath}}}) > t_{\text{obs}} \mid U_{\mathcal{M}} = u_{\text{obs}}, \boldsymbol{Y}_n \in A_{\widehat{\jmath}}\big) \qquad (4)$$

is the uniformly most powerful (UMP) conditional post-selection confidence distribution for θ.

Consequently,

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

(iii) The shortest $100(1-\alpha)\%$ confidence intervals among all coverage proper confidence curves for the pseudo-true value $\theta^*_{M_{\widehat{\jmath}}}$ for every $0 < \alpha < 1$ are obtained by the $\alpha/2$ and

$1-\alpha/2$ *quantiles of* $C_{n,|\widehat{\jmath}}(\theta, \boldsymbol{Y}_n)$, *or equivalently by the* $1-\alpha$ *level set of the confidence curve* $cc_{n,|\widehat{\jmath}} : \Theta \to [0,1] : \theta \mapsto cc_{n,|\widehat{\jmath}}(\theta) = |1 - 2C_{n,|\widehat{\jmath}}(\theta_{M_{\widehat{\jmath}}}, \boldsymbol{Y}_n)|$.

Working with a confidence curve is graphically enlightening to compare the performance of different methods to produce confidence intervals. See for example the left panels in Figure 1. A confidence curve provides confidence intervals at all levels and the confidence intervals' length is depicted by the spread of the curve. The guaranteed confidence level is obtained since for any $\alpha \in (0,1)$, the level set $cc_{n,|\widehat{\jmath}}(\theta) = 1-\alpha$ (a horizontal line at $1-\alpha$ in the graph) consists of two values: $\hat{\theta}_{\text{left}}$ and $\hat{\theta}_{\text{right}}$ for which by definition holds that $C_{n,|\widehat{\jmath}}(\hat{\theta}_{\text{left}}, \boldsymbol{Y}_n) = \alpha/2$ and $C_{n,|\widehat{\jmath}}(\hat{\theta}_{\text{right}}, \boldsymbol{Y}_n) = 1-\alpha/2$, hence $P(\theta \in (\hat{\theta}_{\text{left}}, \hat{\theta}_{\text{right}})) = 1-\alpha$.

Note that the optimal shortest confidence intervals are obtained directly via the confidence distribution. The main advantage is that we do not need to invert pivotal quantities as is typically done in selective inference (e.g. Tibshirani et al., 2018).

3.2 A continuity correction for discrete distributions

The finite sample optimality result in Proposition 1 only applies to continuous distributions. For discrete distributions we may use a continuity correction to approximate the uniformly most powerful conditional post-selection confidence distribution. A 'half'-correction $C_{n,|\widehat{\jmath}}(\theta_{M_{\widehat{\jmath}}}, \boldsymbol{Y}_n) = P(T > t_{\text{obs}} \mid U_{\mathcal{M}} = u_{\text{obs}}, \boldsymbol{Y}_n \in A_{\widehat{\jmath}}) + 1/2\ P(T = t_{\text{obs}} \mid U_{\mathcal{M}} = u_{\text{obs}}, \boldsymbol{Y}_n \in A_{\widehat{\jmath}})$ has been used by Schweder and Hjort (2002) while Veronese and Melilli (2018) proposed the use of the geometric mean of fiducial densities that for exponential families coincide with the asymptotic confidence distributions for the discrete case. The continuity correction for discrete distributions has been studied theoretically by Stone (1969, Appendix 5, Theorem 1). There it was found that using the constant 1/2 has the best properties. We refer to this paper for further details. This choice was further advocated by Schweder and Hjort (2002, 2016). For other approaches for discrete distributions, see Blaker (2000).

An additional approximation is required for binary data with continuous covariates as $U_{\mathcal{M}} = u_{\text{obs}}$ might only be reached by the observed $\boldsymbol{y}_n$, therefore conditioning on $U_{\mathcal{M}} = u_{\text{obs}}$ might imply conditioning on $\boldsymbol{Y}_n$. A solution is to allow some error threshold value δ such that $C_{n,|\widehat{\jmath}}(\theta_{M_{\widehat{\jmath}}}, \boldsymbol{Y}_n) = P(T > t_{\text{obs}} \mid \|U_{\mathcal{M}}\|^2 \leq \|u_{\text{obs}}\|^2 + \delta, \boldsymbol{Y}_n \in A_{\widehat{\jmath}}) + 1/2\ P(T = t_{\text{obs}} \mid \|U_{\mathcal{M}}\|^2 \leq \|u_{\text{obs}}\|^2 + \delta, \boldsymbol{Y}_n \in A_{\widehat{\jmath}})$, with $\|\cdot\|$ the Euclidean norm and δ determined by the magnitude of X and the sample size n. The idea is to set δ as small as possible to have $U_{\mathcal{M}}$ in the neighborhood of u_{obs} but still big enough to have some information left over after conditioning. In practice, one might perform a grid search to choose δ, as discussed in the next section.

3.3 Accounting for misspecification of all models

Even though, our theory requires the selected model to be overspecified, when the true generating process is unknown the selected model might be misspecified. In that case, the sufficient statistics $\widetilde{T}(\boldsymbol{y}_n; X)$ are still sufficient for the pseudo-true parameter vector $(\theta^*_{M_{\widehat{\jmath}}}, \eta^{*\top}_{M_{\widehat{\jmath}}})$. However, as the true distribution of $\boldsymbol{Y}_n$ might not have the natural parametrization form of the exponential family distributions in (3), the true distribution of $\widetilde{T}(\boldsymbol{y}_n; X)$ is unknown and it might not have the optimal properties of the exponential family distributions.

In sections 4.2 and 4.3 we find empirical evidence that the simulated distribution of $T|(U_{\mathcal{M}} = u_{\text{obs}}, \boldsymbol{Y}_n \in A_{\widehat{\jmath}})$ still works for inference on $\theta^*_{M_{\widehat{\jmath}}}$ also when the selected model is misspecified. A theoretical difficulty in the case of completely misspecified models is the lack of information on the distributions of the sufficient statistics.

To account for model misspecification we adjust the Monte-Carlo sampling scheme for the computation of $C_{n,\widehat{\jmath}}(\theta_{M_{\widehat{\jmath}}}, \boldsymbol{y}_n)$ that is conditional on the sufficient statistics (Lindqvist and Taraldsen, 2005; Schweder and Hjort, 2016) and conditional on the selection event. Such scheme was previously used for normal models only. For more details for the Gaussian case, see Garcia-Angulo and Claeskens (2022, Sec. 5).

Step 1. We compute the sufficient statistics for the parameter of interest as well as for all the nuisance parameters in the model set $\mathcal{M}$. For the observed sample $\boldsymbol{y}_n$ and X we compute $(T(\boldsymbol{y}_n; X), U_{\mathcal{M}}^\top(\boldsymbol{y}_n)) = (t_{\text{obs}}, u_{\text{obs}}^\top)$. It is clear that all nuisance parameters are required since the decision to select a model depends on which other models are in $\mathcal{M}$. As an example, think about computing an information criterion such as AIC for all models in $\mathcal{M}$ and selecting the model for which the AIC value is smaller than all other models' AIC values.

Step 2. For a grid of candidate values for the parameter of interest θ in the selected model, we sample from the model distribution under two constraints. First, that the sufficient statistics for the nuisance parameters computed with the new data are the same as the observed values from step 1. Second, that the same model is selected using the sampled data instead of the original data. The generating model uses as parameter values the candidate value for θ and estimated values for η such that the constraints hold.

In practice, we might turn the first constraint into an optimization problem, where for sampling, we find values for η that minimize the Euclidean norm of the difference between the observed sufficient statistics of the nuisance parameters and the sampled ones. Once that first constraint is satisfied, we test whether the selected model is the same in the sampled data. For more details see Garcia-Angulo and Claeskens (2022, Sec. 5). This constrained generation is computationally demanding. For instance, to generate samples for a grid of 20 candidate values in the simulation in Section 4.3, it takes on average 5.9 minutes (sd= 2.4 min, 5 replicates). An algorithm to speed up this step is currently under investigation.

Step 3. A linear interpolation is performed on the set of grid points for θ and the empirical probability that the newly computed sufficient statistics for θ with the generated data exceed the value t_{obs}. This gives the confidence distribution $C_{n,\widehat{\jmath}}(\theta_{M_{\widehat{\jmath}}}, \boldsymbol{y}_n)$. To get the confidence curve we compute $|1 - 2C_{n,\widehat{\jmath}}(\theta, \boldsymbol{y}_n)|$ for a range of values for θ.

The result is referred to as Post-cc1.

For binary response, we might choose δ from a grid search. After step 1, for the smallest (or largest) candidate value for the parameter of interest θ, we sample at least 10 times from the model distribution under the constraint that $\|u^*_{\mathcal{M}} - u_{\text{obs}}\|^2 = \delta$ and the model selection constraint. Here $u^*_{\mathcal{M}}$ is the vector of sufficient statistics for the nuisance parameters calculated using the sampled data. We choose the smallest δ for which the sampled response vectors are different from $\boldsymbol{y}_n$. Once δ is defined, we follow steps 2 and 3.

To account for model misspecification the method Post-cc2 adjusts step 2 in the algorithm which implies a modification of step 3 too. For this correction, the Post-cc2 sampling scheme introduces extra variability by the use of a sandwich standard error estimator of $\hat{\theta}_{M_{\widehat{\jmath}}}$ in the selected model. This estimator is denoted by $\widetilde{\sigma}^2(\theta_{M_{\widehat{\jmath}}})$. For examples of such estimators, see

MacKinnon and White (1985).

Step 2 (modified). For a grid of candidate values for the parameter of interest θ in the selected model, we sample from the model distribution under the same constraints as in step 2. However, for each candidate value ϑ, the generating model uses as parameter values randomly generated values $\vartheta_b \sim N(\vartheta, \widetilde{\sigma}^2(\theta_{M_{\widehat{\jmath}}}))$ for θ and estimated values for η such that the constraints hold.

Step 3 (modified). The linear interpolation of step 3 gives a conservative distribution $\widetilde{C}_{n,|\widehat{\jmath}}(\theta_{M_{\widehat{\jmath}}}, \boldsymbol{Y}_n)$ obtained using the sample from the step 2 (modified). Conservative confidence curves are obtained as $|1 - 2\widetilde{C}_{n,|\widehat{\jmath}}(\theta_{M_{\widehat{\jmath}}}, \boldsymbol{Y}_n)|$.

Indeed, step 2 (modified) corrects for possible misspecification in the data at the price of wider confidence curves.

Example: Poisson regression: computation under misspecification. Assume $\beta = (\beta_1, \beta_2)^\top$ where $\theta = \beta_1$ is of interest and $\eta = \beta_2$ is a nuisance parameter. The model selection set $\mathcal{M}$ consists of two Poisson regression models, M_1 for which $E(Y_i|x_{i1}) = \exp(x_{i1}\beta_1)$ and M_2 for which $E(Y_i|x_{i2}) = \exp(x_{i1}\beta_1 + x_{i2}\beta_2)$. Assume M_2 is selected and it might be misspecified. To simulate a confidence distribution for β_1, first (step 1) we compute the values of the sufficient statistics $t_{\text{obs}} = \sum_{i=1}^n x_{i1}y_i$ and $u_{\text{obs}} = \sum_{i=1}^n x_{i2}y_i$. Then, (step 2 modified) we choose a grid of candidate values for β_1. For each candidate value ϑ in the grid we sample B times data $(y_{1,b} \ldots, y_{n,b})$ for $b = 1, \ldots, B$ from M_2 with a mean specified as $E(Y_i|x_{i2}) = \exp(x_{i1}\vartheta_b + x_{i2}\beta_2^o)$. Here ϑ_b is generated each of the B times from a $N(\vartheta, \widetilde{\sigma}^2(\beta_{1,M_2}))$ where $\widetilde{\sigma}(\beta_{1,M_2})$ is the sandwich standard error estimator of $\hat{\beta}_{1,M_2}$ in the selected model M_2. β_2^o is estimated such that the $u_{\text{obs}} = \sum_{i=1}^n x_{i2}y_{i,b}$. We perform the same model selection procedure we used in the original data now using $(y_{1,b} \ldots, y_{n,b})$ instead and check whether M_2 is selected. We redo this until we have B samples satisfying all constraints. Finally we follow the linear interpolation of step 3 to obtain $\widetilde{C}_{n,|\widehat{\jmath}}(\beta_{1,M_2}, \boldsymbol{Y}_n)$ as in step 3 (modified).

4 Simulation study

We compare both methods Post-cc1 and Post-cc2 to three other methods. 1. A "naive" approach that ignores the model selection step and pretends as if the selected model was given beforehand and is correct. For the naive method confidence distributions are obtained as $C_n(\theta_{M_{\widehat{\jmath}}}, \boldsymbol{Y}_n) = F_{n-p}(\mathcal{T})$ where F_{n-p} is the cumulative distribution function of a t-distribution with $n-p$ degrees of freedom and $\mathcal{T} = (\theta - \hat{\theta}_{M_{\widehat{\jmath}}})/\hat{\sigma}(\theta)_{M_{\widehat{\jmath}}}$ the t-statistic in the selected model.
2. We also compare with Post-AIC confidence intervals (Charkhi and Claeskens, 2018), which were developed especially for inference after selection by AIC and are based on asymptotic results. This method also conditions on the model selected by AIC but is not finite sample exact and leads to conservative inference, meaning too wide confidence intervals.
3. For Simulation 3 we compare to a general method for post selection inference that is unlike the other methods in the comparison not conditional on the selected model. For normal linear models Berk et al. (2013) coined the acronym PoSI (post-selection inference). For logistic regression models we use the version of Bachoc et al. (2020), which is a simultaneous uniform post-selection inference method and is guaranteed to give conservative inference

with confidence intervals having at least the wanted confidence level.

Neither of these post-selection methods was used before in the context of confidence distributions, only single-level confidence intervals were studied in the literature. The simulations show a more complete picture by considering the confidence curves.

4.1 Simulation 1. Poisson regression

The data were generated from a Poisson regression model with log-link such that $E[Y_i] = \exp(\sum_{j=1}^{6} \beta_j x_{i,j})$, for $i = 1, \ldots, 100$. The true value $\beta^\top = (1.2, -0.4, 1.6, -0.05, 0, 0)$. We set $x_{i,1} = 1$ and $(x_{i,2}, \ldots, x_{i,6})^\top \sim N(\mathbf{0}_5, \Omega)$ where Ω is the variance-covariance matrix with an equi-correlation structure set equal to 0.25. The set of models $\mathcal{M}$ consists of 32 models obtained by all possible combinations of the covariates but the intercept which is included in all models. We generated 1000 data sets which select an overparametrized model $M_{\hat{\jmath}}$ with parameters $(\beta_1, \ldots, \beta_5)$. The selection procedure is AIC. To obtain the conservative confidence curves Post-cc2, we use as $\widetilde{\sigma}(\theta_{M_{\hat{\jmath}}})$ in the modified sampling procedure, the estimated standard error in the selected model for $\hat{\beta}_{M_{\hat{\jmath}},r}$, $r = 2, 5$ which in this case is equivalent to the sandwich estimator as the selected model is correct.

Figure 1 summarizes the results of this simulation. We observe that the average width of the curve using our proposed method Post-cc1 is between the naive approach and the Post-AIC method. The average naive confidence curve for β_5 is too narrow leading to drastic under-coverage for $\beta_{M_{\hat{\jmath}},5} = 0$. The naive approach for a true relatively big non-zero parameter such as $\beta_{M_{\hat{\jmath}},2} = -0.4$ gives valid inference. On the other hand, PostAIC and Post-cc2 are conservative methods, for instance, both methods show over-coverage for $\beta_{M_{\hat{\jmath}},2}$, even though, Post-AIC is much more conservative. However, for $\beta_{M_{\hat{\jmath}},5}$, Post-AIC shows under-coverage for $1-\alpha$ confidence intervals up to level 0.5, while Post-cc2 still shows over-coverage. The problem of conservative methods is that the uniformity requirement of a confidence distribution is clearly violated. Even though Post-AIC and Post-cc2 can be used for conservative inference, they may not satisfy the second requirement of a post-selection confidence distribution, that is, the distribution evaluated at its pseudo-true parameter value might not follow a uniform distribution.

4.2 Simulation 2. Poisson regression with heteroscedasticity in the data

We now show what happens if all 32 models in $\mathcal{M}$ are misspecified. To this end, we redo the previous simulation in Section 4.1 but this time the data were generated from a negative binomial regression model with log-link $P(Y_i = y) = \binom{y+1.3-1}{y}\left(\frac{\mu_i}{\mu_i+1.3}\right)^y\left(\frac{1.3}{\mu_i+1.3}\right)^{1.3}$ with $\mu_i = \exp(\sum_{r=1}^{6} \beta_r x_{i,r})$, $i = 1, \ldots, 100$. The set $\mathcal{M}$ consists of Poisson regression models with all possible combinations of predictors. We keep the selection procedure and the selected model $M_{\hat{\jmath}}$ as in Section 4.1. For the estimated standard errors we use a heteroscedasticity consistent covariance matrix estimator in the selected model known as HC3 proposed by MacKinnon and White (1985). HC3 is asymptotically equivalent to the classical White's sandwich estimator but has better finite sample properties and in simulation studies it shows better performance for sample sizes smaller than 250 (see, Long and Ervin, 2000).

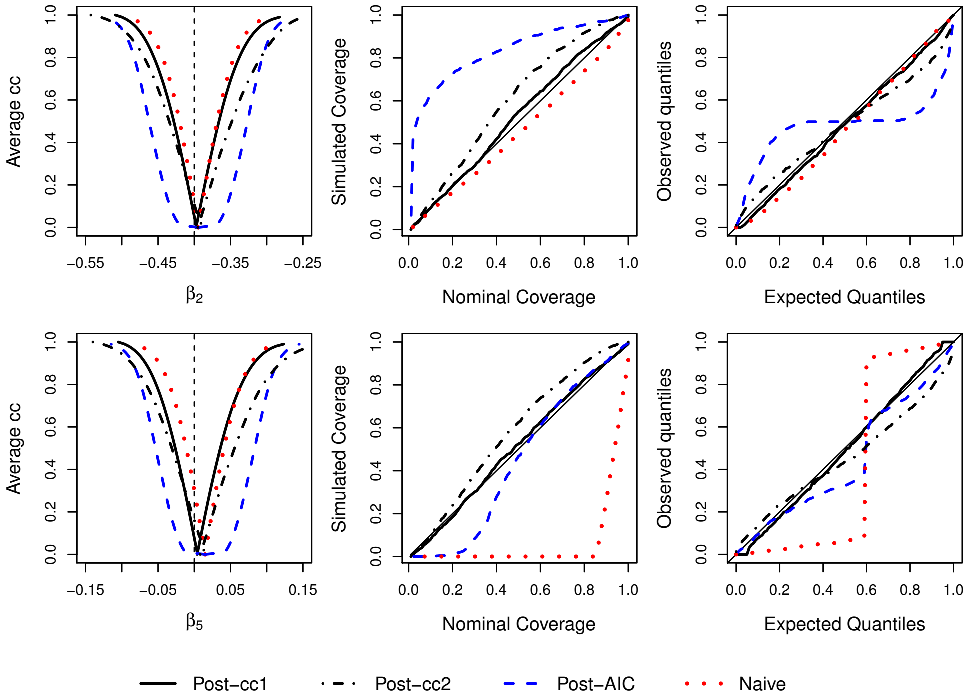

Figure 1: Left: Average confidence curves over 1000 replications for the Poisson regression parameters β_2 and β_5 when the selected model by AIC is correctly specified. The true parameter values are indicated with a dashed vertical line. Center: Simulated mean coverage of the $1-\alpha$ confidence intervals with $\alpha = [0,1]$, for $\beta_{M_{\hat{\jmath}},2} = -0.4$ and $\beta_{M_{\hat{\jmath}},5} = 0$. Right: Quantiles of the simulated distribution of $C_{n|\hat{\jmath}}(-0.4, \boldsymbol{Y}_n)$ and $C_{n|\hat{\jmath}}(0, \boldsymbol{Y}_n)$ for β_2 and β_5, respectively, versus expected quantiles of a $\mathcal{U}[0,1]$. The naive method fails for β_5, while Post-AIC is too conservative for β_2 and has the widest intervals for β_5. Post-cc1 is uniformly most powerful and indeed works best for both parameters. Using Post-cc2, when in doubt of possible misspecification, we lose power resulting in wider confidence intervals with slight overcoverage.

For the naive approach we compute two t-statistics, first using the estimated standard error for the coefficient of interest in the selected model and next using the sandwich estimator HC3. For Post-AIC confidence intervals we use the HC3 estimated covariance matrix in the full model as in the simulation study of Charkhi and Claeskens (2018). Once again, we summarize the results for β_2 and β_5 in Figure 2, showing that the HC3 estimated covariance matrix well captures the heteroscedasticity in the data. The simulated coverage using our proposed method Post-cc2 is proper while keeping the average width of the confidence curves almost the same as for the naive curves. The naive intervals using the HC3 estimated covariance matrix give proper coverage for $\beta_{M_{\hat{\jmath}},2} = -0.4$ but fail for $\beta_{M_{\hat{\jmath}},5} = 0$. Post-AIC confidence curves are conservative for all the parameters, as expected.

4.3 Simulation 3. Logistic regression

We study the approximate optimal conditional post-selection confidence distributions for a parameter of interest on a selected logistic model under two scenarios, when it is correct and

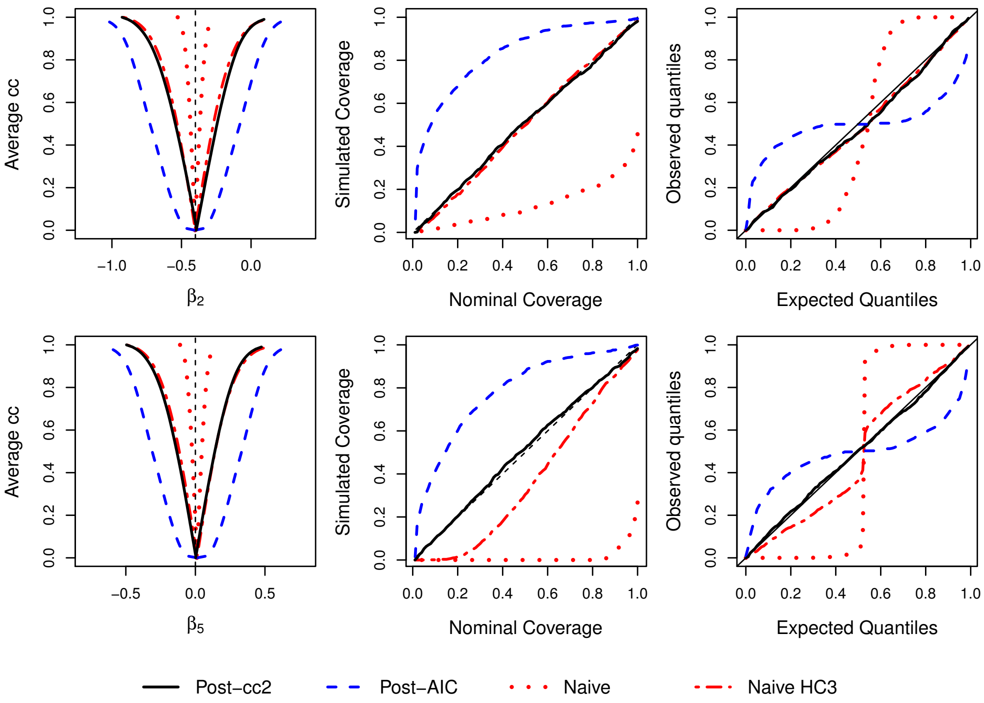

Figure 2: Left: Average confidence curves over 1000 replications for the regression parameters β_2 and β_5 when the data are generated using a negative binomial model and $\mathcal{M}$ contains only Poisson regression models. The selected model by AIC is therefore misspecified. The true parameter values are indicated with a dashed vertical line. Center: Simulated mean coverage of the $1-\alpha$ confidence intervals with $\alpha = [0, 1]$, for $\beta_{M_{\hat{\jmath}},2} = -0.4$ and $\beta_{M_{\hat{\jmath}},5} = 0$. Right: Quantiles of the simulated distribution of $C_{n|\hat{\jmath}}(-0.4, \mathbf{Y}_n)$ and $C_{n|\hat{\jmath}}(0, \mathbf{Y}_n)$ for β_2 and β_5, respectively, versus quantiles of a $\mathcal{U}[0, 1]$. Post-cc2 gives the correct coverage with narrow intervals, while the naive methods have undercoverage and Post-AIC has overcoverage with wider intervals.

when it is misspecified. In the first setting, the data were generated from a logistic regression model $Y_i \sim \text{Bernoulli}(p_i)$ with $\text{logit}(p_i) = \sum_{j=1}^{5} \beta_j x_{i,j}$, for $i = 1, \ldots, 30$. The true values for the parameters are $\beta^\top = (0.1, 2, 0.1, 0, 0)$. We set $x_{i,1} = 1$ and $(x_{i,2}, \ldots, x_{i,5})^\top \sim N(\mathbf{0}_4, \Omega_4)$ where Ω_4 is the variance-covariance matrix with an equi-correlation structure set equal to 0.25.

In the second setting, the true generating process is $Y_i \sim \text{Bernoulli}\{\Phi(\sum_{j=1}^{6} \beta_j x_{i,j})\}$, with Φ the cdf of a standard normal distribution, which corresponds to a probit model. The true values for the parameters are $\beta^\top = (0.1, 2, 0.1, 0, 0, -0.3)$. We set $x_{i,1} = 1$ and the covariates $(x_{i,2}, \ldots, x_{i,6})^\top \sim N(\mathbf{0}_5, \Omega_5)$ with Ω_5 having the same equi-correlation structure set equal to 0.25. We assume that we fail to observe the covariate $x_{i,6}$, therefore the parameter β_6 is not estimated by any of the candidate models.

All working candidate models in $\mathcal{M}$ are logistic regression models in both settings. The selection procedure is as follows. We start with a full logistic model $Y_i \sim \text{Bernoulli}(p_i)$ with p_i as specified above, for $i = 1, \ldots, 30$, and perform a backward selection based on t-tests. In the first step, we compute four t-statistics $\mathcal{T}_{M,r} = \hat{\Sigma}_{M,r}^{-1/2}\ \hat{\beta}_{M,r}$ with $\hat{\Sigma}_{M,r}^{-1/2}$ the (r, r) element of the estimated covariance matrix of $\hat{\beta}_M$, for $r = 2, \dots, 5$ and discard the covariate with the smallest $|\mathcal{T}_{M,r}|$ as long as it is smaller than the critical value for $t_{0.05/2,30-5}$. Here $t_{\alpha/2,n-|M|}$ is the $1 - \alpha/2$ quantile of a t-distribution with $n - |M|$ degrees of freedom. We repeat the first step under the reduced model once and discard another covariate adjusting the degrees of freedom of the critical value with $|M| = 4$. After this, the final model $M_{\hat{\jmath}}$ contains only two covariates plus an intercept. We generate 1000 data sets such that after this selection procedure the final selected model contains the parameters $(\beta_1, \beta_2, \beta_3)$ and all $|\mathcal{T}_{M_{\hat{\jmath}},r}| > t_{0.05/2,30-3}$, for $r = 1, 2, 3$. This procedure is part of a "significance hunting" strategy.

We are interested in the effect of β_3 whose true value is relatively small for the sample size $n = 30$. For the first setting, the true value of interest is $\beta_3 = 0.1$ as the selected model is correct. For the second setting, the pseudo-true value $\beta^*_{M_{\hat{\jmath}}}$ is the solution to the equation $\sum_{i=1}^{n} \{b'(x_i^\top \beta^*_{M_{\hat{\jmath}}}) - b'(\zeta_i)\} x_i = 0$, with $x_i = (x_{i,1}, \dots, x_{i,5})^\top$, $\zeta_i = \Phi(\sum_{j=1}^{6} \beta_j x_{i,j})$ and $b(\cdot) = \log\{1 + exp(\cdot)\}$. The pseudo-true value of interest is the third component $\beta^*_{M_{\hat{\jmath}},3}$. As the covariates are randomly generated in each of the 1000 data sets, the pseudo-true value is different each time. Its average value in our study is 0.092 with standard deviation 0.134.

For the approximation in the conditioning $U_{\mathcal{M}}$ we use $\delta = 3$. Figure 3 illustrates the empirical evidence of this study in terms of average confidence curves, simulated mean coverage of confidence intervals and qq-plots of the simulated distribution at the pseudo-true value compared to a uniform distribution. The upper panel corresponds to the first setting, where $M_{\hat{\jmath}}$ is correct and the lower panel to the second setting where $M_{\hat{\jmath}}$ is misspecified. Due to misspecification we use a sandwich covariance matrix estimator HC3 to all methods displayed. Post-cc1 seems approximately correct when the selected model is correct while ignoring the selection step, the naive method is biased and overoptimistic for the true relatively small $\beta_3 = 0.1$. When the selected model is misspecified Post-cc1 slightly corrects the bias of the naive method on average but produces narrower confidence curves and fails on coverage as expected. A method correcting for misspecification as Post-cc2 is needed. Post-cc2 leads to valid conservative inference for the best approximating value $\beta^*_{M_{\hat{\jmath}},3}$. In both settings, Post-cc2 produces tighter confidence curves than the PoSI curves. As PoSI is valid for any selection method, it is expected to be more conservative than the conditional approach for a specific selection procedure.

5 Application: The levee failure data

As an application to another type of generalized linear model, we construct post-selection confidence curves for a logistic regression model that has been analysed before by naive methods. The data were collected and analyzed by Flor et al. (2010) and we retrieved it from the University of Florida repository. As the authors specified, the goal of the analysis was to test the relative importance of geologic, geomorphic, and other physical factors that have led to levee failures in the Mississippi River. The data set has 70 observations in the Middle Mississippi River collected over 120 years. The response variable is 1 if there was a levee failure and 0 otherwise. There are 11 covariates, Channel fills: presence (1) or absence (0) of channel fills at the site of levee failure, Borrow pits: presence (1) or absence (0) of borrow

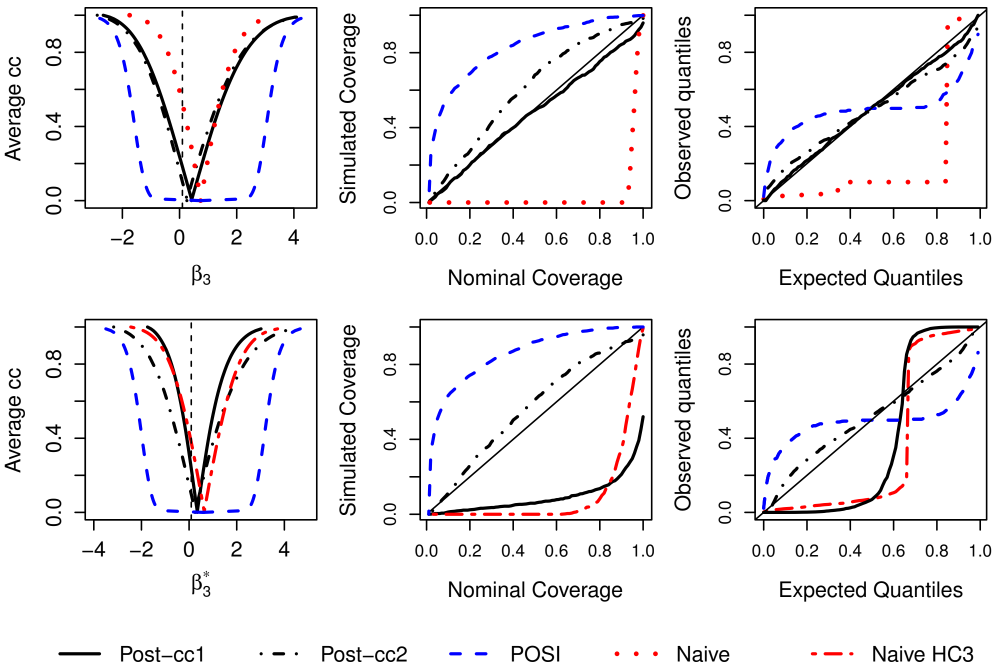

Figure 3: Left: Average confidence curves over 1000 replications for the regression parameters β_3 when the data are generated using a logistic (up) and a probit (bottom) model and $\mathcal{M}$ contains only logistic regression models. The model is selected by a backward selection procedure based on t-tests. The true and 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 $\beta_{M_{\hat{\jmath}},3} = 0.1$ and mean $\bar{\beta}^*_{M_{\hat{\jmath}},3} = 0.092$. Right: Quantiles of the simulated distribution of $C_{n|\hat{\jmath}}(0.1, \boldsymbol{Y}_n)$ (up) and $C_{n|\hat{\jmath}}(\beta^*_{M_{\hat{\jmath}},3}, \boldsymbol{Y}_n)$ (bottom), versus quantiles of a $\mathcal{U}[0, 1]$. Post-cc2 still produces good coverage and narrower average confidence curves than PoSI even when the model is misspecified.

pits, Meander: 4 levels categorical variable for location on a meander sequence (1: inside bend, 2: outside bend, 3: chute, 4: straight), Channel's width: width of channel in meters, Floodway's width: width of floodway in meters, Constriction factor: ratio for constriction factor, Land cover: 4 levels categorical variable for land cover type (1: Open water, 2: Grass, 3: Agricultural, 4: Forest), Vegetative: width of vegetative buffer, Sinuosity: a ratio of the channel's sinuosity, Dredging: a ratio of dredging intensity, Revetment: presence (1) or absence (0) of bank revetment. See Flor et al. (2010) for more details about the data. As the continuous variables have different scales, we standardize them for this analysis.

We redo the selection procedure applied by the authors in the original analysis. They used two different selected models, the first one named "conservative" selects the covariate if and only if it has an individual p-value smaller than 5% when regressing only that covariate against the response. For the second selected model, they relax the threshold for selection to a p-value smaller than 20% and they name it a "liberal" model. This "informal" model selection procedure can be categorized within significance hunting practice by a one at a time simple regression model. The "conservative" selected model has only one predictor and the

		point estimate	p-value
Intercept	Naive	-0.894	0.024
	Post-cc2	-0.774	0.354
Channel fills	Naive	1.587	0.002
	Post-cc2	1.215	0.209

Table 2: "Conservative" selected model for levee failure data. Point estimates and p-values ignoring (naive) and including (post-cc2) model selection by significance hunting at 5%.

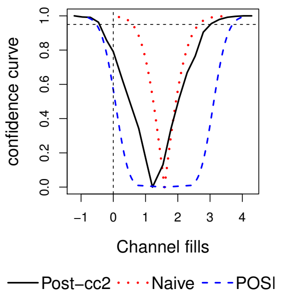

Figure 4: Levee failure data. Confidence curves for parameter in the "conservative" model ignoring (naive) and including (post-cc2 and POSI) model selection.

"liberal" selected model has 5 predictors. We construct the naive confidence curves which coincide with the analysis performed by the authors for the inference on the selected model without any multiple testing correction and compare them with the post-selection confidence curves (Post-cc2) for the parameters on the selected model and the POSI curves (See Figures 4 and 5). We allow for possible heteroscedasticity in the data and use the HC3 sandwich variance estimator and we smoothed the obtained approximated $\widetilde{C}_{n,|\widehat{\jmath}}(\beta_{M_{\widehat{\jmath}},r}, \boldsymbol{Y}_n)$ using linear interpolation with the R-function `approx`. We also obtain a new point estimate defined as the median of $\widetilde{C}_{n,|\widehat{\jmath}}(\beta_{M_{\widehat{\jmath}},r}, \boldsymbol{Y}_n)$ and a post-selection p-value defined as $2\min\{\widetilde{C}_{n,|\widehat{\jmath}}(0, \boldsymbol{Y}_n), 1 - \widetilde{C}_{n,|\widehat{\jmath}}(0, \boldsymbol{Y}_n)\}$ for testing $H_0 : \beta_{M_{\widehat{\jmath}},r} = 0$ versus $H_a : \beta_{M_{\widehat{\jmath}},r} \neq 0$ (See Tables 2 and 3).

Using the naive inference, the presence of channel fills seems to be the only significant variable at 5% related to the levee failures in the Middle Mississippi River in both models. However, accounting for model selection, this is no longer the case, as all variables become non-significant at the 5% level. The biggest change after accounting for the effect of model selection is for channel's width variable in the liberal model, whose naive p-value is 0.08 compared to the post-selection p-value of 0.75 and its post-selection confidence curve is shifted and tight close to zero. We observe the same pattern for dredging intensity. Compared to POSI curves, we observe that post-cc produces almost always tighter confidence curves and therefore shorter confidence intervals.

Remark The variable meander sequence has an unbalanced design with 21 observations collected inside of a meander, 20 along chutes, 23 from straight sections and only 6 outside of a meander. The estimated standard error of the outside bend regression parameter is much smaller using the HC3 estimator than the standard estimation. This causes the difference in the width of the confidence curves and it needs to be interpreted carefully. "Land cover type" has a drastic unbalance in its levels with only 3 observations in open water and a big difference in the response variance of this group compared to the others. We decided to keep the same categorization in order to make our results comparable to the original analysis but do not provide inferential analysis for the open water category.

	Naive		Post-cc2	
	point estimate	p-value	point estimate	p-value
Intercept	-1.416	0.032	-0.417	0.000
Channel fills	1.997	0.005	1.384	0.384
Channel's width	-0.654	0.084	0.049	0.904
Land cover: grass	-0.838	0.421	-0.880	0.554
Land cover: agriculture	-0.751	0.253	-0.384	0.676
Minder: inside	1.335	0.093	0.041	0.954
Minder: outside	-0.003	0.999	-0.036	0.992
Minder: chute	0.365	0.630	-0.268	0.708
Dredging	-0.407	0.197	-0.188	0.598

Table 3: "Liberal" selected model for levee failure data: Point estimates and p-values ignoring (naive) and including (post-cc2) model selection by significance hunting at 20% level using one at a time simple regressions.

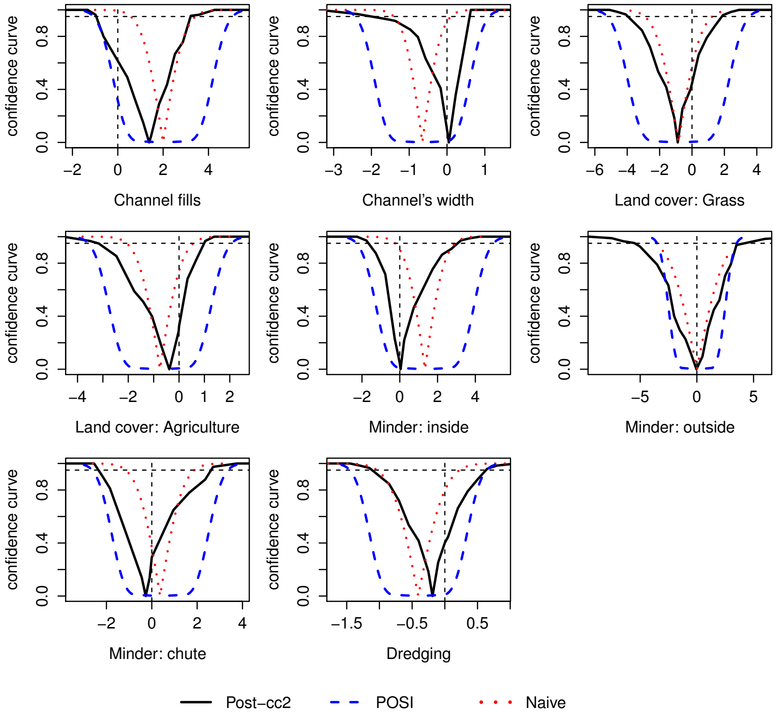

Figure 5: Levee failure data. Confidence curves for parameters in the "Liberal" selected model ignoring (naive) and including (post-cc and POSI) model selection. The vertical dashed line indicates the zero value. The horizontal dashed line is for the endpoints of the 95% confidence intervals.

6 Discussion

Working with confidence distributions has as a major advantage that no separate study is required for hypothesis testing and for the construction of confidence intervals, the latter were mostly studied before for a single confidence level only. This approach yields intervals for all confidence levels. Another major advantage as compared to the available results for selective inference is that no pivotal quantity needs to be inverted. Indeed, confidence intervals are directly obtained as quantiles of the confidence distribution or as level sets of the confidence curve. A graphical representation of the confidence curves allows for an easy visual comparison of different methods, see for example Figure 4.

The finite sample exactness of our results for continuous distributions is another point of distinction from the asymptotic results that are available in the literature. For discrete distributions, the 'half-correction' works fine in practice as the simulations have illustrated.

Confidence distributions for a linear combination of parameters such as for the estimation of the linear model part $x_i^\top \beta$ can be dealt with in a similar way, see Garcia-Angulo and Claeskens (2022, Sec. 4.2).

While the technical limitation required to have in the model set at least one over-parametrized model in order to guarantee the theoretical results for the sufficient statistics to hold within the exponential family, our simulations in misspecified models have illustrated that the method also applies to such settings. An adjustment using a model-robust, or 'sandwich' variance estimator is advised in case of doubt of the model correctness. Conservative results are obtained in such case.

7 Proof of Proposition 1

Proof. Only (i) requires a proof. The other parts follow by properties of the confidence distribution. The proof of this proposition follows the main reasoning as that of Garcia-Angulo and Claeskens (2022) except in this case we assume that both the true generating model and the selected model are in the class of the exponential family, see (3). Define the vector of sufficient statistics for the nuisance parameters in the selected model as $U_{\widehat{\jmath}} = U(\boldsymbol{Y}_n; X_{M_{\widehat{\jmath}}})$. Then, $U_{\mathcal{M}} = \left(U_{\widehat{\jmath}}^\top, (U_{\widehat{\jmath}}^c)^\top\right)^\top$ where $U_{\widehat{\jmath}}^c$ is the vector of sufficient statistics for the nuisance parameters in the other models in $\mathcal{M}$ where duplicated values have been removed. In the classical exponential family theorem, when $\boldsymbol{Y}_n$ is distributed as (3), the conditional distribution of $T|U_{\widehat{\jmath}} = u$ is again an exponential family distribution with one parameter θ (Lehmann and Romano, 2006, Lemma 2.7.2). Moreover, as the parameter space contains an open rectangle in $\mathbb{R}^k$, with $k = |M_{\widehat{\jmath}}|$, $U_{\widehat{\jmath}}$ is a complete sufficient statistic for η. By additionally conditioning on $U_{\widehat{\jmath}}^c$, the distribution of $T|U_{\mathcal{M}}$ remains a one-parameter exponential family. This extra conditioning is needed to fix the selection regions given by the model selection methods which use all the sufficient statistics in $\mathcal{M}$. By doing so, the event $\boldsymbol{Y}_n \in A_{\widehat{\jmath}}$ implies that the domain of $T|U_{\mathcal{M}}$ is restricted to fixed parts in $\mathbb{R}$. Suppose, without loss of generality that $\mathrm{dom}(T|U_{\mathcal{M}} = u_{\mathrm{obs}}, \boldsymbol{Y}_n \in A_{\widehat{\jmath}}) = \{t = t(y; X_{M_{\widehat{\jmath}}}) \in \mathbb{R} : a \leq t \leq b\}$, with fixed a and b given by the specificities of the selection method. Then, $T|U_{\mathcal{M}}$ follows a truncated exponential family distribution with a single parameter θ. The rest of the proof follows as in the proof of Proposition 1 of Garcia-Angulo and Claeskens (2022), where given the strictly increasing likelihood ratio property of the exponential families, $C_{n,|\widehat{\jmath}}(\theta, \boldsymbol{y}_n) = P(T(\boldsymbol{Y}_n; X_{M_{\widehat{\jmath}}}) > t_{\mathrm{obs}} \mid U_{\mathcal{M}} = u_{\mathrm{obs}}, \boldsymbol{Y}_n \in A_{\widehat{\jmath}})$ has the optimal properties of confidence distributions based on sufficient statistics. □

Acknowledgements

The authors thank the reviewers for their helpful comments and questions. Funding is acknowledged from the Research Foundation Flanders and from 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

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

Belloni, A., Chernozhukov, V., and Wei, Y. (2016). Post-selection inference for generalized linear models with many controls. *Journal of Business & Economic Statistics*, 34(4):606–619.

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

Blaker, H. (2000). Confidence curves and improved exact confidence intervals for discrete distributions. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 28(4):783–798.

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

Fithian, W., Sun, D., and Taylor, J. (2014). Optimal inference after model selection. Arxiv 1410.2597.

Flor, A., Pinter, N., and Remo, J. W. F. (2010). Evaluating levee failure susceptibility on the Mississippi River using logistic regression analysis. *Engineering Geology*, 116(1):139 – 148.

Garcia-Angulo, A. and Claeskens, G. (2022). Exact uniformly most powerful post-selection confidence distributions. *Scandinavian Journal of Statistics*, page to appear.

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

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

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

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.

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.

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.

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.

Stone, M. (1969). The role of significance testing: Some data with a message. *Biometrika*, 56(3):485–493.

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. 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. *The 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.

Veronese, P. and Melilli, E. (2018). Fiducial, confidence and objective Bayesian posterior distributions for a multidimensional parameter. *Journal of Statistical Planning and Inference*, 195:153–173.

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.