On various confidence intervals post-model-selection

Leeb, Hannes and Pötscher, Benedikt M. and Ewald, Karl

University of Vienna, University of Vienna, University of Technology Vienna

2014

On various confidence intervals post-model-selection

Hannes Leeb[1], Benedikt M. Pötscher[1], and Karl Ewald[2]

[1]University of Vienna
[2]Vienna University of Technology

January 2014

Abstract

We compare several confidence intervals after model selection in the setting recently studied by Berk et al. (2013), where the goal is to cover not the true parameter but a certain non-standard quantity of interest that depends on the selected model. In particular, we compare the PoSI-intervals that are proposed in that reference with the 'naive' confidence interval, which is constructed as if the selected model were correct and fixed a-priori (thus ignoring the presence of model selection). Overall, we find that the actual coverage probabilities of all these intervals deviate only moderately from the desired nominal coverage probability. This finding is in stark contrast to several papers in the existing literature, where the goal is to cover the true parameter.

1 Introduction and Overview

There is ample evidence in the literature that model selection can have a detrimental impact on subsequently constructed inference procedures like confidence sets, if these are constructed in the 'naive' way where the presence of model selection is ignored. Such results are reported, for example, by Brown (1967); Buehler and Feddersen (1963); Dijkstra and Veldkamp (1988); Kabaila (1998, 2009); Kabaila and Leeb (2006); Leeb (2006); Leeb and Pötscher (2003, 2005, 2006a,b, 2008a,b); Olshen (1973); Pötscher (1991, 2006); Pötscher and Leeb (2009); Pötscher and Schneider (2009, 2010, 2011); Sen (1979); Sen and Saleh (1987).

Recently, Berk, Brown, Buja, Zhang, and Zhao (2013) proposed a new class of confidence intervals, so-called PoSI-intervals, which correct for the presence of model selection, in the sense that these intervals guarantee a user-specified minimal coverage probability, even if the model has been selected in a data-driven way. However, the setting of Berk et al. (2013) differs from earlier studies, because they consider confidence intervals for *a different quantity of interest*: In the aforementioned analyses, the quantity of interest (the coverage target) is always a fixed parameter or sub-parameter of the data-generating model. Berk et al. (2013), on the other hand, consider a different and non-standard coverage target that depends on the selected model. [Even if an overall correct model is assumed, that non-standard coverage target does not coincide with a parameter in the model, except for degenerate and trivial situations.] By design, the PoSI-intervals do not provide a solution to the more traditional problem, where the goal is to cover a parameter in the overall model after model selection.

Berk et al. (2013) motivate the need for PoSI-intervals by the poor performance of the 'naive' interval as observed in the studies mentioned in the first paragraph of this section. However, these studies do not cover the performance of the 'naive' procedures post-model-selection when the coverage target is as in Berk et al. (2013). This raises the question of how the 'naive' interval performs when it is used to cover the coverage target considered in Berk et al. (2013). The main contribution of this paper is to answer this. In particular, we compare 'naive' confidence intervals and PoSI-intervals in the setting of Berk et al. (2013). [The results in the present paper are partly based on Ewald (2012), and we refer to this thesis for additional results and discussion.]

We find that the minimal coverage probability of the 'naive' interval is slightly below the nominal one, while that of the PoSI interval is slightly above, when the coverage target is as in Berk et al. (2013) and when AIC or similar procedures are used for model selection. In the scenarios that we consider, the coverage probabilities of all these intervals are within 5% of the nominal coverage probability, the only exception being one scenario that is designed specifically so that the difference between these intervals is most pronounced (design 3 in Section 4). In the more traditional setting where the coverage target is a parameter in the overall model, however, all these intervals generally fail to deliver the desired minimal coverage probability. [Note that the PoSI-interval is not designed to deal with this coverage target.] For illustration, consider the scenario depicted by the solid curves in Figure 1 on page 9: There, a 'naive' confidence interval post-model-selection with nominal coverage probability 0.95 has a minimal coverage probability of about 0.91 and the corresponding PoSI-interval has a minimal coverage probability of about 0.96, if the coverage target is as in Berk et al. (2013). But if the coverage target is a parameter in the overall model, the minimal coverage probabilities of the 'naive' interval and of the PoSI-interval drop to about 0.56 and 0.62, respectively.

The paper is organized as follows: In Section 2, we introduce the data-generating process, the model-selection procedures, the coverage targets, and various confidence procedures including the PoSI-intervals. We consider the same assumptions and constructions as Berk et al. (2013). The (minimal) coverage probabilities of 'naive' intervals and of PoSI-intervals are studied in Section 3 and Section 4. In particular, Section 3 contains an explicit finite-sample analysis of these procedures in a simple scenario with two nested candidate models. Section 4 contains a simulation study where we compare these intervals in three more complex scenarios; the first scenario is also studied by Kabaila and Leeb (2006), and the other two scenarios are taken from Berk et al. (2013). Finally, in the Appendix, we present an example with a coverage target that is similar to, but slightly different from, those considered in Berk et al. (2013). The interesting feature of this example is that the 'naive' confidence interval here is valid, in the sense that its coverage probability is never below the nominal level.

2 Coverage Targets and Confidence Intervals

Throughout, we consider a set of n homoskedastic Gaussian observations with mean vector $\mu \in \mathbb{R}^n$ and common variance $\sigma^2 > 0$, i.e.,

$$y \quad = \quad \mu + u, \tag{2.1}$$

where $u \sim N(0, \sigma^2 I_n)$. We further assume that we have an estimator $\hat{\sigma}^2$ for σ^2 that is independent of all the least-squares estimators that will be introduced shortly. For the estimator $\hat{\sigma}^2$, we either assume that is distributed as a chi-squared random variable with r degrees of freedom multiplied by σ^2/r, i.e., $\hat{\sigma}^2 \sim \sigma^2\chi^2_r/r$, for some $r \geq 1$; or we assume that the variance is known a-priori, in which case we set $\hat{\sigma}^2 = \sigma^2$ and $r = \infty$. Unless noted otherwise, all considerations that follow apply to both the known-variance case and the unknown-variance case. The joint distribution of y and $\hat{\sigma}$ depends on the unknown parameters $\mu \in \mathbb{R}^n$ and $\sigma > 0$, and will be denoted by $\mathbb{P}_{\mu,\sigma}$.

For the available explanatory variables, consider a fixed $n \times p$ matrix X, where we allow for $p > n$. We consider models where y is regressed on a (non-empty) subset of the regressors in X. For each model $M \subseteq \{1, \ldots, p\}$ with $M \neq \emptyset$, write X_M for the matrix of those columns of X whose indices lie in M. Writing M as $M = \{j_1, \ldots, j_{|M|}\} \subseteq \{1, \ldots, p\}$, we thus have $X_M = (X_{j_1}, \ldots, X_{j_{|M|}})$, where X_j denotes the j-th column of X, and where $|M|$ denotes the size of M. Write $\mathcal{M}$ for the collection of all candidate models under consideration. Throughout, we only consider submodels of full column rank, i.e., we assume that the rank of X_M equals $|M|$ and satisfies $1 \leq |M| \leq n$ for each $M \in \mathcal{M}$.

Under a candidate model $M \in \mathcal{M}$, y is modeled as

$$y \quad = \quad X_M\beta_M + v_M,$$

where β_M corresponds to the orthogonal projection of μ from (2.1) onto the column-space of X_M, i.e., $\beta_M = (X_M'X_M)^{-1}X_M'\mu$. The least-squares estimator corresponding to the model M will be denoted by $\hat{\beta}_M$, i.e., $\hat{\beta}_M = (X_M'X_M)^{-1}X_M'y$. The working model M is correct if $X_M\beta_M = \mu$; in that case, we have $v_M = u$. Otherwise, i.e., if $X_M\beta_M \neq \mu$, the working model is incorrect, and we have $v_M = \mu - X_M\beta_M + u$. Irrespective of whether the working model is correct, we always have $\hat{\beta}_M \sim N(\beta_M, \sigma^2(X_M'X_M)^{-1})$; in particular, $\hat{\beta}_M$ is an unbiased estimator for β_M, irrespective of whether or not the model M is correct. As noted earlier, we assume that the variance estimator $\hat{\sigma}^2$ is independent of the estimators $\hat{\beta}_M$ for $M \in \mathcal{M}$.

To identify the regression coefficient of a given regressor X_j in a model M it appears in, we write $\beta_{j\cdot M}$ for that component of β_M that corresponds to the regressor X_j for each $j \in M$. Similarly, the components of $\hat{\beta}_M$ are indexed as

$\hat{\beta}_{j\cdot M}$ for $j \in M$. This convention is called 'full model indexing' in Berk et al. (2013).

Consider now a model selection procedure, i.e., a data-driven rule that selects a model $\hat{M} \in \mathcal{M}$ from the pool of candidate models, and the resulting post-model-selection estimator $\hat{\beta}_{\hat{M}}$. The coverage target considered in Berk et al. (2013) is $\beta_{\hat{M}}$, or components thereof. Note that this coverage target is random, because it depends on the outcome of the model selection procedure.

Remark 2.1. (i) At least one author of the present paper believes that the interpretation of this coverage target is debatable: For example, the meaning of the first coefficient of $\beta_{\hat{M}}$ depends on the selected model and hence also on the training data; the same applies to the dimension of $\beta_{\hat{M}}$. We refer to Berk et al. (2013) for further discussion and motivation for studying $\beta_{\hat{M}}$.

(ii) While the model (2.1) is non-parametric, the distributional requirements on $\hat{\sigma}^2$ are rather restrictive. However, these requirements are fulfilled if (2.1) is replaced by the *parametric* model $y = X\beta + u$, if X is assumed to be of full column rank $p < n$, if $\hat{\sigma}^2$ is the usual unbiased variance estimator in that model, and if r is set to $n-p$. In that case, the true parameter β in the overall model is well-defined and will then typically be the prime target of statistical inference.

In this paper, we will mainly focus on confidence intervals for the coefficient of one particular predictor in the selected model. Without loss of generality, assume that X_1 is the predictor of interest, and that the coverage target is $\beta_{1\cdot\hat{M}}$. To ensure that this quantity is always well-defined, we assume that the first predictor X_1 is contained in all candidate models under consideration, i.e., we assume that $1 \in M$ for each $M \in \mathcal{M}$. We seek to construct confidence intervals for $\beta_{1\cdot\hat{M}}$ that are of the form

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

for some constant $K > 0$, with $\hat{\sigma}^2_{1\cdot M}$ defined by $\hat{\sigma}^2_{1\cdot M} = \hat{\sigma}^2[(X_M'X_M)^{-1}]_{1,1}$, where $[\ldots]_{1,1}$ denotes the first diagonal element of the indicated matrix. For a given level $1-\alpha$ with $0 < \alpha < 1$, the constant K should be chosen such that the minimal coverage probability is at least $1-\alpha$, i.e., such that

$$\inf_{\mu,\sigma} \mathbb{P}_{\mu,\sigma}\left(\beta_{1\cdot\hat{M}} \in \hat{\beta}_{1\cdot\hat{M}} \pm K\hat{\sigma}_{1\cdot\hat{M}} \right) \geq 1-\alpha. \tag{2.2}$$

Because the distribution of $(\hat{\beta}_{1\cdot M} - \beta_{1\cdot M})/\hat{\sigma}_{1\cdot M}$ is independent of unknown parameters and also independent of M, it follows, for fixed M, that a confidence interval for $\beta_{1\cdot M}$ with minimal coverage probability $1-\alpha$ is given by the textbook interval $\hat{\beta}_{1\cdot M} \pm K_N\hat{\sigma}_{1\cdot M}$, where K_N is the $(1-\alpha/2)$-quantile of the distribution of $(\hat{\beta}_{1\cdot M} - \beta_{1\cdot M})/\hat{\sigma}_{1\cdot M}$ – a standard normal distribution in the known-variance case and a t-distribution with r degrees of freedom in the unknown-variance case. In view of this, it is tempting to consider, as a confidence interval for $\beta_{1\cdot\hat{M}}$, the interval $\hat{\beta}_{1\cdot\hat{M}} \pm K_N\hat{\sigma}_{1\cdot\hat{M}}$. Because this construction ignores the model selection step and treats the selected model $\hat{M}$ as fixed, we will call this the 'naive' confidence interval.

The PoSI-interval developed in Berk et al. (2013) is obtained by first constructing simultaneous confidence intervals for the components of β_M that are centered at the corresponding components of $\hat{\beta}_M$, for each $M \in \mathcal{M}$, with coverage probability $1-\alpha$: More formally, the PoSI-constant K_P is such that

$$\inf_{\mu,\sigma} \mathbb{P}_{\mu,\sigma}\left(\beta_{j\cdot M} \in \hat{\beta}_{j\cdot M} \pm K_P \hat{\sigma}_{j\cdot M} :\ j \in M, M \in \mathcal{M}\right) \quad = \quad 1-\alpha, \qquad (2.3)$$

where the quantities $\hat{\sigma}^2_{j\cdot M}$ are defined like $\hat{\sigma}^2_{1\cdot M}$ but with j replacing 1. By construction, the PoSI-constant K_P is such that we obtain simultaneous confidence intervals for the components of $\beta_{\hat{M}}$ that are centered at the corresponding components of $\hat{\beta}_{\hat{M}}$. In other words, we have

$$\inf_{\mu,\sigma} \mathbb{P}_{\mu,\sigma}\left(\beta_{j\cdot \hat{M}} \in \hat{\beta}_{j\cdot \hat{M}} \pm K_P \hat{\sigma}_{j\cdot \hat{M}} :\ j \in \hat{M}\right) \quad \geq \quad 1-\alpha. \qquad (2.4)$$

In particular, (2.2) holds when K_P replaces K. For computing the constant K_P, we note that the probability in (2.3) can also be written as $\mathbb{P}(|\hat{\beta}_{j\cdot M} - \beta_{j\cdot M}|/\hat{\sigma}_{j\cdot M} \leq K_P :\ j \in M, M \in \mathcal{M})$. This probability is not hard to compute, because it involves only the random variables $(\hat{\beta}_{j\cdot M} - \beta_{j\cdot M})/\hat{\sigma}_{j\cdot M}$, which are (correlated) standard normal in the known-variance case and (correlated) t-distributed in the unknown variance case. In particular, the probability in (2.3) does not depend on μ or σ^2. Similar considerations apply, mutatis mutandis, to the constant K_{P1} that is introduced in the following paragraph.

A modification of this procedure, which is also proposed in Berk et al. (2013), is useful when inference is focused on a particular component of $\beta_{\hat{M}}$, instead of on all components. Recall that the coverage target in (2.2) is the first component of $\beta_{\hat{M}}$, i.e., $\beta_{1\cdot\hat{M}}$. The PoSI1-constant K_{P1} provides simultaneous confidence intervals for $\beta_{1\cdot M}$ centered at $\hat{\beta}_{1\cdot M}$ for each $M \in \mathcal{M}$. In particular, K_{P1} is chosen so that

$$\inf_{\mu,\sigma} \mathbb{P}_{\mu,\sigma}\left(\beta_{1\cdot M} \in \hat{\beta}_{1\cdot M} \pm K_{P1} \hat{\sigma}_{1\cdot M} :\ M \in \mathcal{M}\right) \quad = \quad 1-\alpha.$$

Again by construction, (2.2) holds when K_{P1} replaces K.

Like the PoSI-constants discussed so far, other procedures for controlling the family-wise error rate can be used. Consider, for example, Scheffé's method: Recall that X denotes the matrix of all available explanatory variables, and note that $(\hat{\beta}_{j\cdot M} - \beta_{j\cdot M})$ is a linear function of $Y-\mu$, i.e., a function of the form $\nu'(Y-\mu)$, for a certain vector ν in the span of X. The Scheffé constant K_S is chosen such that

$$\mathbb{P}_{\mu,\sigma}\left(\sup_{\substack{\nu\in\mathrm{span}(X)\\ \nu\neq 0}} \frac{\nu'(Y-\mu)}{\hat{\sigma}\|\nu\|} \leq K_S\right) \quad = \quad 1-\alpha.$$

Then the relations (2.3), (2.4), and, in particular, (2.2) hold when K_S replaces both K and K_P. Note that the probability in the preceding display does not depend on μ and σ, and that the constant K_S is easily computed as follows: Let p denote the rank of X. In the known-variance case, K_S is the square root of the $(1-\alpha)$-quantile of a chi-square distribution with p degrees of freedom. In the unknown-variance case, K_S is the square root of the product of p and the $(1-\alpha)$-quantile of an F-distribution with p and r degrees of freedom.

Using the constants K_P, K_{P1} or K_S gives valid confidence intervals post-model-selection, i.e., intervals that satisfy (2.2), because these constants give simultaneous confidence intervals for all quantities of interest that can occur; for example, (2.4) follows from (2.3), which in turn guarantees that (2.2) holds when K_P replaces K. One advantage of this is that the minimal coverage probability is guaranteed, *irrespective* of the model selection procedure $\hat{M}$. In particular, coverage is guaranteed even if the model is selected by statistically inane methods like the SPAR-procedure mentioned in Section 4.9 of Berk et al. (2013). The price for this is that the PoSI constants K_P and K_{P1} may be overly conservative for a *particular* model selection procedure $\hat{M}$.

Lastly, we will also consider the obvious approach where one chooses the smallest constant K such that (2.2) is satisfied. We will denote this constant by K_*. This is, of course, a well-known standard construction; see Bickel and Doksum (1977, p.170) for example. By definition, the interval in (2.2) with K_* replacing K is the shortest interval of that form whose minimal coverage probability is $1-\alpha$. Note that K_* depends on the model selection procedure in question, and that computation of this quantity can be cumbersome as it requires computation of the finite-sample distribution of $\hat{\beta}_{1\cdot\hat{M}}/\hat{\sigma}_{1\cdot\hat{M}}$. However, explicit computation of this constant is feasible in some cases (cf. the results in Section 3 and also the more general results of Leeb and Pötscher (2003)), and this constant can also be computed or approximated in a variety of other scenarios (for example, by adapting the results of Pötscher and Schneider (2010) or the procedures of Andrews and Guggenberger (2009)). Also note that we have $K_* \leq K_{P1} \leq K_P \leq K_S$ by construction.

The procedures discussed so far are concerned with coverage targets like $\beta_{\hat{M}}$ that depend on the selected model. This should be compared to the more classical parametric setting where the coverage target is the underlying true parameter: Assume that the data is generated by a linear overall model, i.e., assume that the parameter μ in (2.1) satisfies $\mu = X\beta$ for some overall regressor matrix X. And assume that inference is focused on (components of) the parameter β. In this setting, the effect of model selection on subsequently constructed confidence intervals can be dramatic. For example, Kabaila and Leeb (2006) show that the minimal coverage probability of the 'naive' confidence interval for β_1, i.e., the quantity

$$\inf_{\beta,\sigma} \mathbb{P}_{X\beta,\sigma}\left(\beta_1 \in \hat{\beta}_{1\cdot\hat{M}} \pm K_N \hat{\sigma}_{1\cdot\hat{M}} \right),$$

can be much smaller than the nominal coverage probability $1-\alpha$; in fact, this minimal coverage probability can, e.g., be smaller than 0.5, depending on the regressor matrix X in the overall model $y = X\beta + u$. The main reason for this more dramatic effect is that $\hat{\beta}_{1\cdot M}$ is a biased estimator for β_1 whenever the model M is incorrect, whereas $\hat{\beta}_{1\cdot M}$ is always unbiased for $\beta_{1\cdot M}$. Of course, valid confidence intervals post-model-selection can also be constructed when the coverage target is β_1, namely by replacing K_N in the preceding display by the smallest constant K such that the resulting minimal coverage probability equals $1-\alpha$. For the computation or approximations of this constant in particular situations, we refer to the papers cited in the preceding paragraph.

3 Explicit Finite-Sample Results

In this section we give a finite-sample analysis of the confidence intervals discussed so far, where we consider a simple model selection procedure that selects among two nested models using a likelihood-ratio test. More precisely, let X be an $n \times 2$ matrix of rank 2, and assume that $\mathcal{M} = \{M_1, M_2\}$ with $M_1 = \{1\}$ and $M_2 = \{1, 2\}$ throughout this section. For the model-selector, we set $\hat{M} = M_2$ if $|\hat{\beta}_{2\cdot M_2}|/\hat{\sigma}_{2\cdot M_2}$ is larger than C, and $\hat{M} = M_1$ otherwise, where $C > 0$ is a user-specified constant. [Recall that in the known-variance case, we have $\hat{\sigma}_{j\cdot M} = \sigma^2[(X_M' X_M)^{-1}]_{j,j}$.] Arguably, any reasonable model selection procedure in this setting must be equivalent to a likelihood-ratio test, at least asymptotically; cf. Kabaila and Leeb (2006). In the numerical examples that follow, we will choose $C = \sqrt{2}$, such that the resulting model selector $\hat{M}$ corresponds to selection by the classical Akaike information criterion (AIC). Throughout this section, let $\phi(\cdot)$ and $\Phi(\cdot)$ denote the density and the cumulative distribution function of the univariate standard Gaussian distribution, and set $\Delta(x, c) = \Phi(x + c) - \Phi(x - c)$. And, lastly, we will write ρ for the correlation coefficient between the two components of $\hat{\beta}_{M_2}$, i.e., $\rho = -[(X_{M_2}' X_{M_2})^{-1}]_{1,2}([(X_{M_2}' X_{M_2})^{-1}]_{1,1}[(X_{M_2}' X_{M_2})^{-1}]_{2,2})^{-1/2}$.

The following result describes the coverage probability of the interval $\hat{\beta}_{1\cdot\hat{M}} \pm K\hat{\sigma}_{1\cdot\hat{M}}$ in two scenarios, namely when the coverage target is $\beta_{1\cdot\hat{M}}$ and when the coverage target is $\beta_{1\cdot M_2}$. If the model M_2 is correct, i.e., if we have $\mu = X\beta$ for some $\beta \in \mathbb{R}^2$, and hence also $y = X\beta + u$, then this second scenario reduces to the classical parametric setting described at the end of Section 2; in particular, we then have $\beta_{M_2} = \beta$.

Proposition 3.1. *In the setting of this section, we have*

$$\mathbb{P}_{\mu,\sigma}\left(\beta_{1\cdot\hat{M}} \in \hat{\beta}_{1\cdot\hat{M}} \pm K\hat{\sigma}_{1\cdot\hat{M}}\right) = \\ \mathbb{E}\left[\Delta\left(0, \frac{\hat{\sigma}}{\sigma}K\right)\Delta\left(\zeta, \frac{\hat{\sigma}}{\sigma}C\right) + \int_{-\frac{\hat{\sigma}}{\sigma}K}^{\frac{\hat{\sigma}}{\sigma}K}\left(1 - \Delta\left(\frac{\zeta + \rho z}{\sqrt{1-\rho^2}}, \frac{\frac{\hat{\sigma}}{\sigma}C}{\sqrt{1-\rho^2}}\right)\right)\phi(z)dz\right],$$

and

$$\mathbb{P}_{\mu,\sigma}\left(\beta_{1\cdot M_2} \in \hat{\beta}_{1\cdot\hat{M}} \pm K\hat{\sigma}_{1\cdot\hat{M}}\right) = \mathbb{P}_{\mu,\sigma}\left(\beta_{1\cdot\hat{M}} \in \hat{\beta}_{1\cdot\hat{M}} \pm K\hat{\sigma}_{1\cdot\hat{M}}\right) + \mathbb{E}\left[\left(\Delta\left(\frac{\rho\zeta}{\sqrt{1-\rho^2}}, \frac{\hat{\sigma}}{\sigma}K\right) - \Delta\left(0, \frac{\hat{\sigma}}{\sigma}K\right)\right)\Delta\left(\zeta, \frac{\hat{\sigma}}{\sigma}C\right)\right],$$

with $\zeta = \beta_{2\cdot M_2}/SD(\hat{\beta}_{2\cdot M_2})$, where $SD(\cdot)$ denotes the standard deviation. The expectations on the right-hand sides are taken with respect to $\hat{\sigma}/\sigma$. In the known-variance case, $\hat{\sigma}/\sigma$ is constant equal to one and the expectations are trivial; in the unknown-variance case, $\hat{\sigma}/\sigma$ is distributed like the square root of a chi-squared distributed random variable with r degrees of freedom divided by r, i.e., $\hat{\sigma}/\sigma \sim \sqrt{\chi_r^2/r}$.

Proof. The statements for the known-variance case are simple adaptations of the finite-sample statements of Proposition 3 in Kabaila and Leeb (2006). For the unknown-variance case, it suffices to note that $\hat{\sigma}/\sigma$ is independent of $(\hat{\beta}_{M_1}, \hat{\beta}_{M_2})$. With this, the statements are then obtained by conditioning on $\hat{\sigma}/\sigma$, and by using the formulae for the known-variance case derived earlier. □

Proposition 3.1 provides explicit formulas that also allow us to compute (minimal) coverage probabilities numerically. For the following discussion, fix the values of C and K, i.e., the critical value C of the hypothesis test that is used for model selection, and the value K that governs the length of the confidence interval post-model-selection. We first note that $\mathbb{P}_{\mu,\sigma}(\beta_{1\cdot M_2} \in \hat{\beta}_{1\cdot\hat{M}} \pm K\hat{\sigma}_{1\cdot\hat{M}})$ is strictly smaller than $\mathbb{P}_{\mu,\sigma}(\beta_{1\cdot\hat{M}} \in \hat{\beta}_{1\cdot\hat{M}} \pm K\hat{\sigma}_{1\cdot\hat{M}})$ whenever $\rho\zeta \neq 0$, because the two probabilities differ by a correction factor (namely the expectation term on the right-hand side of the second display in Proposition 3.1) which is negative whenever $\rho\zeta \neq 0$. If $\rho\zeta = 0$, the two probabilities are equal. And if $\rho = 0$, both probabilities are equal to $\Delta(0, K) = \Phi(K) - \Phi(-K)$, irrespective of ζ, as is easily seen. Next, we note that the coverage probabilities depend only on r, ζ and ρ. [Recall that r denotes the degrees of freedom of $\hat{\sigma}^2$ in the unknown-variance, and that we have set $r = \infty$ in the known-variance case.] Note that ζ is a function of the regressor matrix X_{M_2} and of the unknown parameters μ and σ^2, while ρ is a function of X_{M_2} only. Moreover, it is easy to see that the coverage probabilities are symmetric both in ζ and in ρ around the origin. Concerning the influence of r, it can be shown that the coverage probabilities for the known-variance case provide a uniform approximation to those in the unknown variance case, uniformly in the unknown parameters, where the approximation error goes to zero as $r \to \infty$; this follows from the results of Leeb and Pötscher (2003) using standard arguments. In the examples that follow, we found that the results for the known-variance case and for the unknown-variance case are similar, and that these results are visually hard to distinguish from each other, unless r is extremely small like, e.g., 3. We therefore focus on the known-variance case in the following, because it provides a good approximation to the unknown variance case as long as r is not too small.

We proceed to comparing the case where the coverage target is $\beta_{1\cdot\hat{M}}$ as in Berk et al. (2013) with the case where the coverage target is the parameter $\beta_{1\cdot M_2}$, in terms of the coverage probabilities of confidence intervals post-model-selection. For several of the confidence intervals introduced in the preceding section, the results are visualized in Figure 1, for the case where the coverage target is $\beta_{1\cdot\hat{M}}$ (top panel), and for the case where the coverage target is $\beta_{1\cdot M_2}$ (bottom panel). Note that the range of the vertical axes (displaying coverage probability) in the two panels is quite different.

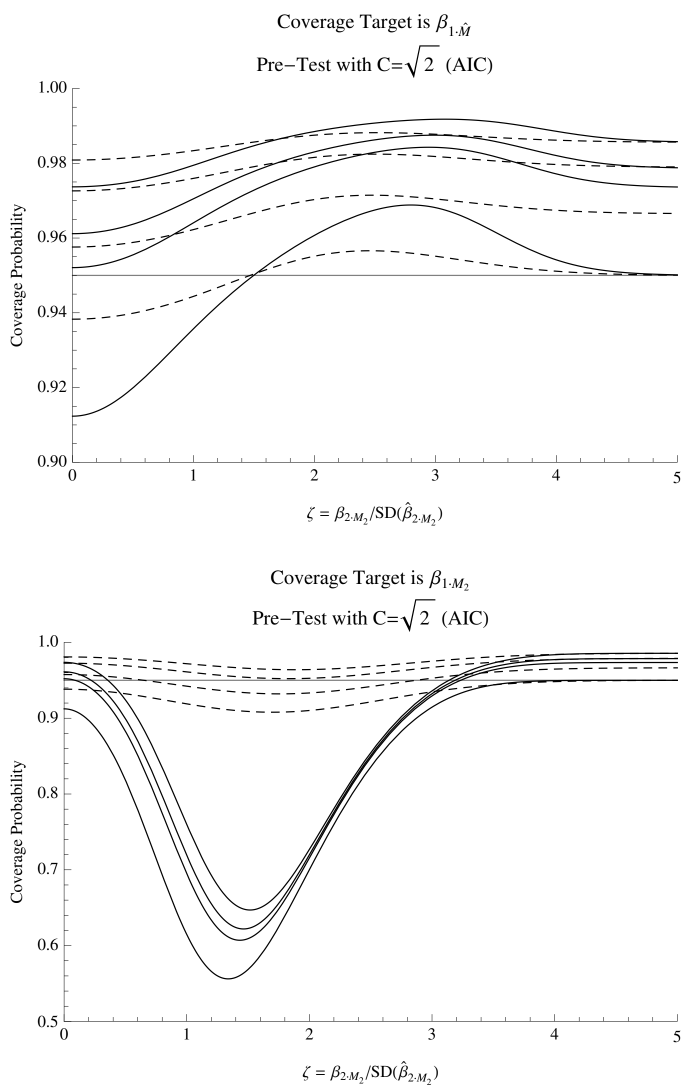

Figure 1: Coverage probability of several confidence intervals in the known-variance case, as a function of the scaled parameter

$\zeta = \beta_{2\cdot M_2}/SD(\hat{\beta}_{2\cdot M_2})$, using the model selection procedure with $C = \sqrt{2}$, i.e., AIC. The nominal coverage probability is $1-\alpha = 0.95$, indicated by a gray horizontal line. The coverage target is $\beta_{1\cdot \hat{M}}$ (top panel) and $\beta_{1\cdot M_2}$ (bottom panel). In each panel, the four solid curves are computed for $\rho = 0.9$, and the four dashed curves are for $\rho = 0.5$. The curves in each group of four are ordered: Starting from the top, the curves show the coverage probabilities for K_S (Scheffé), K_P (PoSI), K_{P1} (PoSI1), and K_N (naive).

In each panel of Figure 1, we see that the effect of model selection on the resulting coverage probabilities depends on the correlation coefficient ρ, with larger values of ρ corresponding to smaller minimal coverage probabilities. But the strength of the effect varies greatly with the scenario, i.e., on whether the coverage target is $\beta_{1\cdot \hat{M}}$ or $\beta_{1\cdot M_2}$. When the coverage target is $\beta_{1\cdot \hat{M}}$ (top panel in Figure 1), we see that the effect of model selection is comparatively minor: The smallest coverage probabilities are always obtained for the 'naive' interval, whose coverage probability here can be smaller as well as larger than the nominal 0.95. Irrespective of the true parameters, the actual coverage probability of the 'naive' interval is quite close to the nominal one here. The other intervals, i.e., the PoSI1- the PoSI- and the Scheffé-interval, all have coverage probabilities larger than 0.95. [The minimal coverage probabilities here are obtained for $\zeta = 0$, but we found this not to be the case for other model selection procedures, i.e., for other values of C.] When the coverage target is $\beta_{1\cdot M_2}$ (bottom panel in Figure 1), however, we get a very different picture: For $\rho = 0.9$, the minimal coverage probability of all the intervals considered there is smaller than 0.95, with minima between 0.55 ('naive') and 0.65 (Scheffé). For $\rho = 0.5$, the minimal coverage probabilities of the 'naive' interval and of the PoSI1-interval are below, while those of the other intervals are above, the nominal 0.95. For very small values of ρ, the coverage probabilities of all the intervals considered in Figure 1 are visually indistinguishable from straight lines as a function of ζ (and hence are not shown here), irrespective of the coverage target. For $\rho = 0.1$, for example, the coverage probability of the 'naive' interval is about 0.95, while that of the other intervals is above 0.95, ordered by their length.

Figure 1 illustrates that the coverage probability of confidence intervals post-model-selection depends crucially on whether the coverage target is $\beta_{1\cdot \hat{M}}$ as in Berk et al. (2013) or the more classical coverage target $\beta_{1\cdot M_2}$. We stress here again that the PoSI-intervals and the Scheffé-interval are not designed to deal with the case where the coverage target is $\beta_{1\cdot M_2}$, and that the performance of these intervals is shown in the bottom panel of Figure 1 only for illustration. For a more detailed analysis of the case where the coverage target is $\beta_{1\cdot M_2}$, we refer to Kabaila and Leeb (2006). In the rest of this section, we focus on the case where the coverage target is $\beta_{1\cdot \hat{M}}$.

We next compare the confidence intervals for $\beta_{1\cdot \hat{M}}$ introduced in Section 2 through their minimal coverage probability as a function of the correlation coefficient ρ. In particular, we compute the quantity (2.2) for various choices of K, namely for K_N ('naive'), for K_P (PoSI), for K_{P1} (PoSI1), for K_S (Scheffé), and for K_* (the smallest valid K). By construction, we have $K_* \leq K_{P1} \leq K_P \leq K_S$, so that the resulting curves of minimal coverage probabilities are also arranged in increasing order.

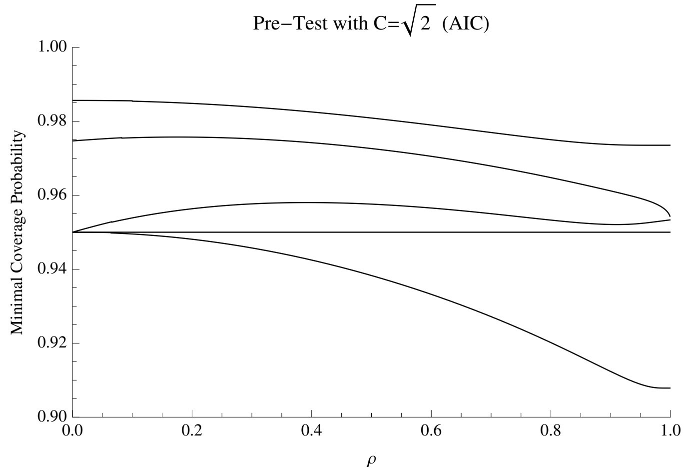

Figure 2: Minimal coverage probabilities of the confidence intervals for $\beta_{1\cdot\hat{M}}$ as a function of ρ in the known-variance case, using the model selection procedure with $C = \sqrt{2}$, i.e., AIC. The nominal coverage probability is $1-\alpha = 0.95$. The curves are ordered: Starting from the top, the curves correspond to the intervals with K_S, K_P, K_{P1}, K_*, and K_N.

All the minimal coverage probabilities shown in Figure 2 are within 5% of the nominal level 0.95. For the 'naive' interval corresponding to K_N, the minimal coverage probability is below 0.95 (except for the trivial case where $\rho = 0$), but not by much. The interval with K_* has a minimal coverage probability of exactly 0.95 by construction. And, again by construction, all other intervals are slightly too large in the sense that their minimal coverage probability exceeds the nominal level 0.95. Overall, the difference between these intervals is not dramatic.

Lastly, we compare the confidence intervals for $\beta_{1\cdot\hat{M}}$ through the values of the constants K in (2.2) that correspond to the intervals in question. By construction, K_S and K_N are constant as a function of ρ. Note that the constants K_N, K_P, K_{P1}, and K_S do not depend on the model selection procedure that is being used, while the constant K_* does depend on $\hat{M}$. For a given model selector $\hat{M}$, the constant K_* is the smallest number K for which (2.2) holds; in particular, the interval corresponding to K has minimal coverage probability smaller/equal/larger than $1 - \alpha$ if and only if K is smaller/equal/larger than K_*.

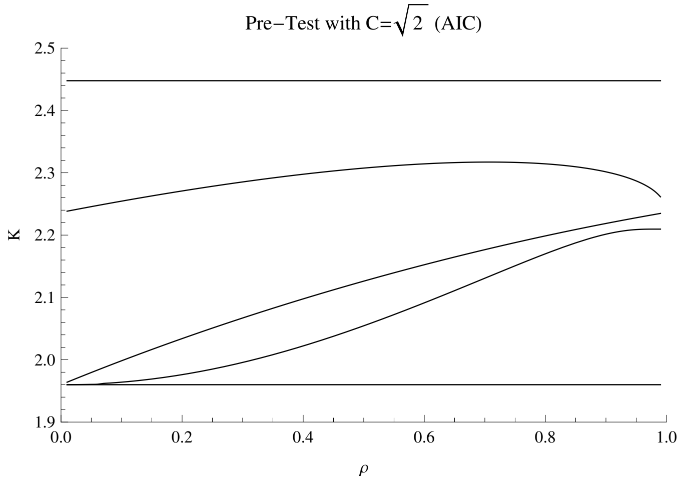

Figure 3: The constants K that govern the width of the confidence intervals as a function of ρ in the known-variance case, using the model selection procedure with $C = \sqrt{2}$, i.e., AIC. The nominal coverage probability is $1 - \alpha = 0.95$. Starting from the top, the five curves show K_S, K_P, K_{P1}, K_*, and K_N.

The interpretation of Figure 3 is similar to that of Figure 2, the main difference being that the lengths considered here are somewhat more distorted than the minimal coverage probabilities considered earlier. The 'naive' interval is up to about 10% too short, while the intervals corresponding to K_{P1}, K_P, and K_S are too long, namely by up to about 5%, 15%, 25%, respectively.

4 Simulation study

We now compare the 'naive' confidence interval and the PoSI-confidence interval for $\beta_{1\cdot\hat{M}}$ by their respective minimal coverage probabilities in a simulation study where the data is generated from an Gaussian linear overall model of the form $Y = X\beta + u$ with 30 observations, 10 explanatory variables, and i.i.d. standard normal errors. Moreover, we also study these intervals when the coverage target is β_1 (instead of $\beta_{1\cdot\hat{M}}$). For the estimator $\hat{\sigma}^2$, we use the usual unbiased variance estimator obtained by fitting the overall model; hence, we have $r = n - p = 20$ here. For the model selector $\hat{M}$, we use the `step()` function in R with its default setting; this corresponds to minimizing the AIC objective function through a greedy general-to-specific search over the 2^{29} candidate models (the regressor of interest, i.e., the first one, is included in all candidate models). Three designs are considered for the design matrix X: For design 1, we take the regressor matrix from the data-example from Section 3 of Kabaila and Leeb (2006) (for which the minimal coverage probability of a 'naive' nominal 95% interval for

β_1 was found to be no more that 0.63 in that paper[1]). For design 2 and 3, respectively, we consider the exchangeable design and the equicorrelated design studied in Sections 6.1 and 6.2 of Berk et al. (2013). The exchangeable design is such that the corresponding PoSI-constant K_P is small asymptotically, and the equicorrelated design corresponds to a large PoSI-constant asymptotically; cf. Theorem 6.1 and Theorem 6.2 of Berk et al. (2013). For the equicorrelated design (design 3), the difference between the PoSI-interval and the 'naive' interval is thus expected to be most pronounced.

More precisely, for the first design, we take the regressor matrix from a dataset of Rawlings (1998) (p.179), where the response is peak flow rate from watersheds, and where the explanatory variables are rainfall (inches), which is the predictor of interest here, as well as area of watershed (square miles), area impervious to water (square miles), average slope of watershed (percent), longest stream flow in watershed (thousands of feet), surface absorbency index (0 = complete absorbency; 100 = no absorbency), estimated soil storage capacity (inches of water), infiltration rate of water into soil (inches/hour), time period during which rainfall exceeded 1/4 inch/hour, and a constant term to include an intercept in the model. Logarithms are taken of the response and of all explanatory variables except for the intercept. For the second design, we define $\mathbf{X}^{(p)}(a)$ as in Section 6.1 of Berk et al. (2013) with $p = 30$ and we choose $a = 10$ here, and we set $X = U\mathbf{X}^{(p)}(a)$, where U is a collection of p orthonormal n-vectors obtained by first drawing a set of n i.i.d. standard Gaussian n-vectors and then applying the Gram-Schmidt procedure. And for the third design, we define $\mathbf{X}^{(p)}(c)$ as in Section 6.2 of Berk et al. (2013) such that the primary predictor of interest is the first one, mutatis mutandis, where we choose $c = \sqrt{0.8/(p-1)}$, and we set $X = V\mathbf{X}^{(p)}(c)$, where V is obtained by drawing an independent observation from the same distribution as U before.

For each of the three design matrices, we simulate coverage probabilities under the model $Y = X\beta + u$ for randomly selected values of the parameter β, we identify those β's for which the simulated coverage probability gets small, and we correct for bias as explained in detail shortly. For example, consider the case where the coverage target is β_1 and where the 'naive' confidence interval is used. We first select 10,000 parameters β by drawing i.i.d. samples from a random p-vector b such that Xb follows a standard Gaussian distribution within the column-space of X. For each of these β's, we approximate the corresponding coverage probability by the coverage rate obtained from 100 Monte Carlo samples. In particular, we draw 100 Monte Carlo samples from the overall model using β as the true parameter. For each Monte Carlo sample, we compute the model selector $\hat{M}$ and the resulting 'naive' confidence interval, and we record whether β_1 is covered or not. The 100 recorded results are then averaged, re-

[1] For the 'naive' intervals considered in Kabaila and Leeb (2006), the error variance is always re-estimated in the selected submodel. Here, on the other hand, we always use the variance estimator $\hat{\sigma}^2$ based on the overall model (in order to be consistent with the setting studied by Berk et al. (2013)). In additional simulations, we found that the coverage probability of the 'naive' interval is typically slightly smaller if the error variance is re-estimated as in Kabaila and Leeb (2006).

sulting in a coverage rate that provides an estimator for the coverage probability of the interval if the true parameter is β. After repeating this for each of the 10,000 β's, we compute the resulting smallest coverage rate as an estimator for the minimal coverage probability of the confidence interval. The smallest coverage rate, as an estimator for the smallest coverage probability, is clearly biased downward. To correct for that, we then take those 1,000 parameters β that gave the smallest coverage rates and re-estimate the corresponding coverage probabilities as explained earlier, but now using 1,000 Monte Carlo samples. For that parameter β that gives the smallest coverage rate in this second run, we run the simulation again but now with 500,000 Monte Carlo samples, to get a reliable estimate of its coverage probability. This procedure is also used to evaluate the performance of the PoSI-interval and also in the case where the coverage target is $\beta_{1\cdot\hat{M}}$, mutatis mutandis. Table 1 summarizes the results.

Target	Interval	Design 1 (watershed data)	Design 2 (exchangeable)	Design 3 (equicorrelated)
$\beta_{1\cdot\hat{M}}$	PoSI CI	1.00	1.00	0.99
	Naive CI	0.89	0.92	0.81
β_1	PoSI CI	0.85	0.91	0.83
	Naive CI	0.62	0.82	0.54

Table 1: Smallest coverage probabilities found in MC study for the coverage targets $\beta_{1\cdot\hat{M}}$, and β_1, and for the PoSI-interval and the 'naive' interval with nominal coverage probability 0.95.

The results of the simulation study reinforce the impression already gained in the theoretical analysis in Section 3: When the coverage target is $\beta_{1\cdot\hat{M}}$, the PoSI-interval is somewhat too long and the 'naive' interval is somewhat too short, resulting in moderate over-coverage and under-coverage, respectively. Both over- and under-coverage are somewhat more pronounced than in the simple model studied in Section 3. But when the coverage target is β_1, then the actual coverage probability of both intervals can be far below the nominal level. As expected, the difference between the 'naive' interval and the PoSI-interval is most pronounced for design 3.

Acknowledgments

We thank the authors of Berk, Brown, Buja, Zhang, and Zhao (2013) for providing us with the code to compute the PoSI-constants used in Section 4; the entire "PoSI-group" at the University of Pennsylvania for inspiring discussions during Hannes Leeb's visit; and Francois Bachoc for constructive feedback. Karl Ewald gratefully acknowledges financial support from Deutsche Forschungsgemeinschaft (DFG) grant FOR916, and Hannes Leeb's research is partially supported by FWF grant P26354.

Appendix: Confidence sets under zero-restrictions post-model-selection

Let y and $\hat{\sigma}^2$ be as in Section 2, and consider $\mathcal{M} = \{M_0, M_1\}$, where each of the two candidate models M_i is full-rank. Suppose we are interested in the coefficient of the first regressor X_1, that is present in M_1 but absent in M_0. In the notation introduced in Section 2, we thus have $1 \in M_1$ and $1 \notin M_0$. As the model-dependent coverage target, which we denote by b_M, we consider the coefficient of X_1, which is not restricted under M_1, and which is restricted to zero under M_0. In other words, we have $b_{M_1} = \beta_{1\cdot M_1}$ and $b_{M_0} = 0$. Let $\hat{M}$ be *any* model selection procedure that chooses only between M_0 and M_1. We consider a 'naive' confidence interval that is defined as

$$I_{\hat{M}} \quad = \quad \begin{cases} \hat{\beta}_{1\cdot M_1} \pm k_N \hat{\sigma}_{1\cdot M_1} & \text{if } \hat{M} = M_1 \\ \{0\} & \text{if } \hat{M} = M_0, \end{cases}$$

where k_N is chosen so that $\mathbb{P}_{\mu,\sigma}(\beta_{1\cdot M_1} \in I_{M_1}) = 1 - \alpha$. [The constant k_N is a standard normal quantile in the known-variance case and a t-quantile in the unknown-variance case.] The actual coverage probability of $I_{\hat{M}}$, as a confidence interval for $b_{\hat{M}}$, is at least equal to the nominal coverage probability $1 - \alpha$, because

$$\begin{aligned} &\mathbb{P}_{\mu,\sigma}(b_{\hat{M}} \in I_{\hat{M}}) \\ &\quad = \quad \mathbb{P}_{\mu,\sigma}(\beta_{1\cdot M_1} \in I_{M_1} \text{ and } \hat{M} = M_1) \;+\; \mathbb{P}_{\mu,\sigma}(0 \in \{0\},\, \hat{M} = M_0) \\ &\quad = \quad \mathbb{P}_{\mu,\sigma}(\beta_{1\cdot M_1} \in I_{M_1} \text{ and } \hat{M} = M_1) \;+\; \mathbb{P}_{\mu,\sigma}(\hat{M} \neq M_1) \\ &\quad = \quad \mathbb{P}_{\mu,\sigma}(\beta_{1\cdot M_1} \in I_{M_1} \text{ or } \hat{M} \neq M_1) \quad \geq \quad 1 - \alpha, \end{aligned}$$

where the inequality in the last step holds in view of the choice of k_N.

References

D. W. K. Andrews and P. Guggenberger. Hybrid and size-corrected subsampling methods. *Econometrica*, **77**:721–762, 2009.

R. Berk, L. Brown, A. Buja, K. Zhang, and L. Zhao. Valid post-selection inference. *Ann. Statist.*, **41**:802–837, 2013.

P. J. Bickel and K. A. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics.* Holden-Day, Oakland, 1977.

L. D. Brown. The conditional level of Student's t test. *Ann. Math. Stat.*, **38**:1068–1071, 1967.

R. J. Buehler and A. P. Feddersen. Note on a conditional property of Student's t. *Ann. Math. Stat.*, **34**:1098–1100, 1963.

T. K. Dijkstra and J. H. Veldkamp. Data-driven selection of regressors and the bootstrap. *Lecture Notes in Econom. and Math. Systems*, **307**:17–38, 1988.

K. Ewald. On the influence of model selection on confidence regions for marginal associations in the linear model. Master's thesis, University of Vienna, 2012.

P. Kabaila. Valid confidence intervals in regression after variable selection. *Econometric Theory*, **14**:463–482, 1998.

P. Kabaila. The coverage properties of confidence regions after model selection. *Int. Statist. Rev.*, **77**:405–414, 2009.

P. Kabaila and H. Leeb. On the large-sample minimal coverage probability of confidence intervals after model selection. *J. Amer. Statist. Assoc.*, **101**:619–629, 2006.

H. Leeb. The distribution of a linear predictor after model selection: unconditional finite-sample distributions and asymptotic approximations. *IMS Lecture Notes - Monograph Series*, **49**:291–311, 2006.

H. Leeb and B. M. Pötscher. The finite-sample distribution of post-model-selection estimators, and uniform versus non-uniform approximations. *Econometric Theory*, **19**:100–142, 2003.

H. Leeb and B. M. Pötscher. Model selection and inference: Facts and fiction. *Econometric Theory*, **21**:21–59, 2005.

H. Leeb and B. M. Pötscher. Can one estimate the conditional distribution of post-model-selection estimators? *Ann. Statist.*, **34**:2554–2591, 2006a.

H. Leeb and B. M. Pötscher. Performance limits for estimators of the risk or distribution of shrinkage-type estimators, and some general lower risk-bound results. *Econometric Theory*, **22**:69–97, 2006b.

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

H. Leeb and B. M. Pötscher. Model selection. In T. G. Andersen, R. A. Davis, J.-P. Kreiß, and Th. Mikosch, editors, *Handbook of Financial Time Series*, pages 785–821, New York, NY, 2008b. Springer.

R. A. Olshen. The conditional level of the F-test. *J. Amer. Statist. Assoc.*, **68**:692–698, 1973.

B. M. Pötscher. Effects of model selection on inference. *Econometric Theory*, **7**: 163–185, 1991.

B. M. Pötscher. The distribution of model averaging estimators and an impossibility result regarding its estimation. *IMS Lecture Notes - Monograph Series*, **52**:113–129, 2006.

B. M. Pötscher and H. Leeb. On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding. *J. Multivariate Anal.*, **100**: 2065–2082, 2009.

B. M. Pötscher and U. Schneider. On the distribution of the adaptive LASSO estimator. *J. Statist. Plann. Inference*, **139**:2775–2790, 2009.

B. M. Pötscher and U. Schneider. Confidence sets based on penalized maximum likelihood estimators in Gaussian regression. *Electron. J. Statist.*, **4**:334–360, 2010.

B. M. Pötscher and U. Schneider. Distributional results for thresholding estimators in high-dimensional Gaussian regression models. *Electron. J. Statist.*, **5**:1876–1934, 2011.

J.O. Rawlings. *Applied Regression Analysis: A Research Tool.* Springer Verlag, New York, NY, 1998.

P. K. Sen. Asymptotic properties of maximum likelihood estimators based on conditional specification. *Ann. Statist.*, **7**:1019–1033, 1979.

P. K. Sen and E. A. K. Md. Saleh. On preliminary test and shrinkage M-estimation in linear models. *Ann. Statist.*, **15**:1580–1592, 1987.