Econometric Theory, **21**, 2005, 21–59. Printed in the United States of America.
DOI: 10.1017/S0266466605050036

MODEL SELECTION AND INFERENCE: FACTS AND FICTION

HANNES LEEB
Yale University

BENEDIKT M. PÖTSCHER
University of Vienna

Model selection has an important impact on subsequent inference. Ignoring the model selection step leads to invalid inference. We discuss some intricate aspects of data-driven model selection that do not seem to have been widely appreciated in the literature. We debunk some myths about model selection, in particular the myth that consistent model selection has no effect on subsequent inference asymptotically. We also discuss an "impossibility" result regarding the estimation of the finite-sample distribution of post-model-selection estimators.

1. INTRODUCTION

In this expository article we discuss some of the problems that arise if one tries to conduct statistical inference in the presence of data-driven model selection. The position we hence take is that a (finite) collection of competing models is given, typically submodels obtained from an overall model through parameter restrictions, and that the researcher uses the data to select one of the competing models.[1] The model selection procedure used here can be based on a (multiple) hypothesis testing scheme (e.g., general-to-specific testing, thresholding as in wavelet regression, etc.), on the optimization of a penalized goodness-of-fit criterion (e.g., Akaike information criterion [AIC], Bayesian information criterion [BIC], final prediction error [FPE], minimum description length [MDL], or any of its numerous variants), or on cross-validation methods. The parameters of the selected model are then estimated (e.g., by least squares or maximum likelihood). Estimators resulting from such a two-step procedure are called "post-model-selection estimators," the classical pretest estimators constituting an important example. As an illustration consider regressor selection in a linear model followed by least-squares estimation of the coefficients of the selected regressors. Here the competing models are submodels of an overall linear regression model (of fixed finite dimension), the submodels being given by zero-restrictions on the regression coefficients.

Address correspondence to Benedikt Pötscher, Department of Statistics, University of Vienna, Universitätsstrasse 5, A-1010, Vienna, Austria; e-mail: Benedikt.Poetscher@univie.ac.at

 0266-4666/05 $12.00

In this paper we do not wish to enter into a discussion of whether or not a two-step procedure as described previously can be justified from a purely decision-theoretic point of view (although we touch upon this important question in the discussion of the mean-squared error of post-model-selection estimators in Sections 2.1 and 2.2 and also in Remark 4.1, which follows). We rather take the pragmatic position that such procedures, explicitly acknowledged or not, are prevalent in applied econometric and statistical work and that one needs to look at their true sampling properties and related questions of inference post model selection. Despite the importance of this problem in econometrics and statistics, research on this topic has been neglected for decades, exceptions being the pretest literature as summarized in Judge and Bock (1978) or Giles and Giles (1993), on the one hand, and the contributions regarding distributional properties of post-model-selection estimators by, e.g., Sen (1979), Sen and Saleh (1987), Dijkstra and Veldkamp (1988), and Pötscher (1991), on the other hand.[2] Only in recent years has this area seen an increase in research activity (e.g., Kabaila, 1995, 1998; Pötscher, 1995; Pötscher and Novak, 1998; Ahmed and Basu, 2000; Kapetanios, 2001; Dukić and Peña, 2002; Hjort and Claeskens, 2003; Kabaila and Leeb, 2004; Leeb and Pötscher, 2003a, 2003b, 2004; Leeb, 2003a, 2003b; Nickl, 2003; Danilov and Magnus, 2004).

The aim of this paper is to point to some intricate aspects of data-driven model selection that do not seem to have been widely appreciated in the literature or that seem to be viewed too optimistically. In particular, we demonstrate innate difficulties of data-driven model selection. Despite occasional claims to the contrary, no model selection procedure—implemented on a machine or not—is immune to these difficulties. The main points we want to make and that will be elaborated upon subsequently can be summarized as follows.[3]

1. Regardless of sample size, the model selection step typically has a dramatic effect on the sampling properties of the estimators that can not be ignored. In particular, the sampling properties of post-model-selection estimators are typically significantly different from the nominal distributions that arise if a fixed model is supposed.
2. As a consequence, naive use of inference procedures that do not take into account the model selection step (e.g., using standard t-intervals as if the selected model had been given prior to the statistical analysis) can be highly misleading.
3. An increasingly frequently used argument in the literature is that consistent model selection procedures allow one to employ the standard asymptotic distributions that would apply if no model selection were performed and that thus the effects of consistent model selection on inference can be safely ignored.[4] Unfortunately, at closer inspection this conclusion turns out not to be warranted at all, and relying on it only creates an illusion of conducting valid inference. In the same vein, the effects of procedures

that consistently choose from a finite set of alternatives (e.g., procedures that consistently decide between $I(0)$ and $I(1)$ or consistently select the number of structural breaks, etc.) on subsequent inference can not be ignored safely. Although it is mathematically true that the use of a consistent model selection procedure entails that the (pointwise) asymptotic distributions of the post-model-selection estimators coincide with the asymptotic distributions that would arise if the selected model were treated as fixed a priori (see, e.g., Pötscher, 1991, Lemma 1), this does not justify the aforementioned conclusion (for the reasons already outlined in Pötscher, 1991, Sect. 4, Remark (iii); and further discussed in Kabaila, 1995).[5]

4. More generally, regardless of whether a consistent or a conservative[6] model selection procedure is used, the finite-sample distributions of a post-model-selection estimator are typically *not* uniformly close to the respective (pointwise) asymptotic distributions. Hence, *regardless of sample size* these asymptotic distributions can *not* be safely used to replace the (complicated) finite-sample distributions.
5. The finite-sample distributions of post-model-selection estimators are typically complicated and depend on unknown parameters. Estimation of these finite-sample distributions is "impossible" (even in large samples). No resampling scheme whatsoever can help to alleviate this situation.

To facilitate a detailed analysis of the effects of selecting a model from a collection of competitors we assume in this paper—as already noted earlier—that one of the competing models is capable of correctly describing the data generating process. Of course, it can always be debated whether or not such an assumption leads to a "test-bed" that is relevant for empirical work, but we shall not pursue this debate here (see, e.g., the contribution of Phillips, 2005, in this issue). The important question of the effects of model selection when selecting only from approximate models will be studied elsewhere.

The points listed previously will be exemplified in detail in Section 2 in the context of a very simple linear regression model, although they are valid on a much wider scope. Because of its simplicity, this example is amenable to a small-sample and also to a large-sample analysis, allowing one to easily get insight into the complications that arise with post-model-selection inference; for results in more general frameworks see Pötscher (1991), Leeb and Pötscher (2003a, 2003b, 2004), and Leeb (2003a, 2003b). Consistent model selection procedures are discussed in Section 2.1, whereas Section 2.2 deals with conservative procedures. Section 2.3 is devoted to the question of estimating the finite-sample distribution of post-model-selection estimators. Shrinkage-type estimators such as Lasso-type estimators, Bridge estimators, and the smoothly clipped absolute deviation (SCAD) estimator, etc., are briefly discussed in Section 3. Section 4 contains some remarks, and Section 5 concludes. Some technical results and their proofs are collected in the Appendixes.

https://doi.org/10.1017/S0266466605050036 Published online by Cambridge University Press

2. AN ILLUSTRATIVE EXAMPLE

In the following discussion we shall—for the sake of exposition—use a very simple example to illustrate the issues involved in model selection and inference post model selection. These issues, however, clearly persist also in more complicated situations such as, e.g., nonlinear models, time series models, etc. Consider the linear regression model

$$y_t = \alpha x_{t1} + \beta x_{t2} + \epsilon_t \quad (1 \le t \le n) \tag{1}$$

under the "textbook" assumptions that the errors ϵ_t are independent and identically distributed (i.i.d.) $N(0, \sigma^2)$, $\sigma^2 > 0$, and the nonstochastic $n \times 2$ regressor matrix X has full rank and satisfies $X'X/n \to Q > 0$ as $n \to \infty$. For simplicity, we shall also assume that the error variance σ^2 is known.[7] It will be convenient to write the matrix $\sigma^2(X'X/n)^{-1}$ as

$$\sigma^2(X'X/n)^{-1} = \begin{pmatrix} \sigma_\alpha^2 & \sigma_{\alpha,\beta} \\ \sigma_{\alpha,\beta} & \sigma_\beta^2 \end{pmatrix}.$$

The elements of this matrix depend on sample size n, but we shall suppress this dependence in the notation. The elements of the limit of this matrix will be denoted by $\sigma_{\alpha,\infty}^2$, etc. It will prove useful to define $\rho = \sigma_{\alpha,\beta}/(\sigma_\alpha \sigma_\beta)$, i.e., ρ is the correlation coefficient between the least-squares estimators for α and β in model (1). Its limit will be denoted by ρ_∞.

Suppose now that the parameter of interest is the coefficient α in (1) and that we are undecided whether or not to include the regressor x_{t2} in the model a priori. (The case where a general linear function $A(\alpha, \beta)'$, e.g., a predictor, rather than α is the quantity of interest is quite similar and is briefly discussed in Remark 4.5.) In other words, we have to decide on the basis of the data whether to fit the unrestricted (full) model or the restricted model with $\beta = 0$. We shall denote the two competing candidate models by U and R (for *unrestricted* and *restricted*, respectively). For any given value of the parameter vector (α, β), the most parsimonious true model will be denoted by M_0 and is given by

$$M_0 = \begin{cases} U & \text{if } \beta \neq 0 \\ R & \text{if } \beta = 0. \end{cases}$$

It is important to note that M_0 depends on the unknown parameters (namely, through β). The least-squares estimators for α and β in the unrestricted model will be denoted by $\hat{\alpha}(U)$ and $\hat{\beta}(U)$, respectively. The least-squares estimator for α in the restricted model will be denoted by $\hat{\alpha}(R)$, and we shall set $\hat{\beta}(R) = 0$. We shall decide between the competing models U and R depending on whether the test statistic $|\sqrt{n}\hat{\beta}(U)/\sigma_\beta| > c$ or not, where $c > 0$ is a user-specified cutoff point. That is, we shall use the model $\hat{M} = U$ if $|\sqrt{n}\hat{\beta}(U)/\sigma_\beta| > c$, and we shall work with $\hat{M} = R$ otherwise. This is a traditional pretest procedure based on the likelihood ratio, but it is worth noting that in the simple example dis-

https://doi.org/10.1017/S0266466605050036 Published online by Cambridge University Press

cussed here it coincides exactly with Akaike's minimum AIC rule in case $c = \sqrt{2}$ and with Schwarz's minimum BIC rule if $c = \sqrt{\log n}$. (We note here in passing that there is a close connection between pretest procedures and information criteria in general; see Remark 4.2.) In fact, in the present example it seems that there is little choice with regard to the model selection procedure other than the choice of c, as it is hard to come up with a reasonable model selection procedure that is not based on the likelihood ratio statistic (at least asymptotically). Now that we have defined the model selection procedure $\hat{M}$, the resulting post-model-selection estimator for the parameter of interest α will be denoted by $\tilde{\alpha} = \hat{\alpha}(\hat{M})$; i.e.,

$$\tilde{\alpha} = \hat{\alpha}(R)\mathbf{1}(\hat{M} = R) + \hat{\alpha}(U)\mathbf{1}(\hat{M} = U).$$

The following simple observations will be useful: The finite-sample distribution of $\tilde{\alpha}$ is a convex combination of the conditional distributions, where the conditioning is on the outcome of the model selection procedure $\hat{M}$:

$$\begin{aligned} P_{n,\alpha,\beta}(\tilde{\alpha} \le t) &= P_{n,\alpha,\beta}(\tilde{\alpha} \le t | \hat{M} = R) P_{n,\alpha,\beta}(\hat{M} = R) \\ &\quad + P_{n,\alpha,\beta}(\tilde{\alpha} \le t | \hat{M} = U) P_{n,\alpha,\beta}(\hat{M} = U) \\ &= P_{n,\alpha,\beta}(\hat{\alpha}(R) \le t | \hat{M} = R) P_{n,\alpha,\beta}(\hat{M} = R) \\ &\quad + P_{n,\alpha,\beta}(\hat{\alpha}(U) \le t | \hat{M} = U) P_{n,\alpha,\beta}(\hat{M} = U), \end{aligned} \tag{2}$$

where $P_{n,\alpha,\beta}$ denotes the probability measure corresponding to the true parameters α, β and sample size n. The model selection probabilities $P_{n,\alpha,\beta}(\hat{M} = U)$ and $P_{n,\alpha,\beta}(\hat{M} = R) = 1 - P_{n,\alpha,\beta}(\hat{M} = U)$ can be evaluated easily and are given by

$$P_{n,\alpha,\beta}(\hat{M} = U) = 1 - (\Phi(c - \sqrt{n}\beta/\sigma_\beta) - \Phi(-c - \sqrt{n}\beta/\sigma_\beta)), \tag{3}$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function (c.d.f.). Cf. Leeb and Pötscher (2003a, Sect. 3.1) and Leeb (2003b, Sect. 3.1).

The subsequent discussion is cast in terms of consistent versus conservative model selection procedures, because this is entrenched terminology.[8] However, despite this terminology, one should not lose sight of the fact that we are given only one sample of *fixed* sample size n together with a *fixed* model selection procedure (e.g., a particular value of the cutoff point c in the present example) and we are interested in the finite-sample properties of this procedure. Any given model selection procedure can now equally well be embedded as a member into a sequence of consistent model selection procedures or into a sequence of conservative procedures for the purpose of asymptotic analysis (by appropriately defining the model selection procedures at the other—fictitious—sample sizes). Of course, the finite-sample properties of the given model selection procedure are unaffected by our choice of the embedding asymptotic framework. Hence, when talking about consistent or conservative sequences of

https://doi.org/10.1017/S0266466605050036 Published online by Cambridge University Press

model selection procedures we are in fact not talking about different procedures but rather about different asymptotic frameworks and their comparative (dis)advantages in revealing the finite-sample properties of a given procedure.

2.1. The Consistent Model Selection Framework

As mentioned in the introduction, proceeding with inference post model selection "as usual" (i.e., as if the selected model were given a priori) is often defended by the argument that a consistent model selection procedure has been used and hence asymptotically the selected model would coincide with the most parsimonious true model, supposedly allowing one to use the standard asymptotic results that apply in case of an a priori fixed model. We now look more closely at the merit of such an argument.

We assume in this section that the cutoff point c in the definition of the model selection procedure $\hat{M}$ is chosen to depend on sample size n such that $c \to \infty$ and $c/\sqrt{n} \to 0$ as $n \to \infty$. Then it is well known (see Bauer, Pötscher, and Hackl, 1988; and also Remark 4.3) that the model selection procedure is a consistent procedure in the sense that

$$P_{n,\alpha,\beta}(\hat{M} = M_0) \xrightarrow{n\to\infty} 1 \tag{4}$$

holds for every α, β; i.e., the probability of revealing the most parsimonious true model tends to unity as sample size increases. Because the event $\{\hat{M} = M_0\}$ is clearly contained in the event $\{\tilde{\alpha} = \hat{\alpha}(M_0)\}$, the consistency property expressed in (4) moreover immediately entails that

$$P_{n,\alpha,\beta}(\tilde{\alpha} = \hat{\alpha}(M_0)) \xrightarrow{n\to\infty} 1 \tag{5}$$

holds for every α, β, where $\hat{\alpha}(M_0)$ denotes the least-squares estimator in the most parsimonious true model. Although this latter "estimator" is infeasible as it makes use of the unknown information whether or not $\beta = 0$, relation (5) shows that the post-model-selection estimator $\tilde{\alpha}$ is a feasible version in the sense that both estimators coincide with probability tending to unity as sample size increases. An immediate consequence of (5) is that the (pointwise) asymptotic distributions of $\tilde{\alpha}$ and $\hat{\alpha}(M_0)$ are identical, regardless of whether $M_0 = U$ or $M_0 = R$. This latter property, which is sometimes called the "oracle" property (Fan and Li, 2001), obviously holds for post-model-selection estimators obtained through consistent model selection procedures in general; cf. Pötscher (1991, Lemma 1) for a formal statement.[9]

So far the preceding discussion seems to support the argument that proceeding "as usual" with inference post consistent model selection is justified. In particular, it seems to suggest that the usual construction of confidence sets remains valid post consistent model selection. Furthermore, observe that (5) entails that the post-model-selection estimator $\tilde{\alpha}$ is asymptotically normally dis-

https://doi.org/10.1017/S0266466605050036 Published online by Cambridge University Press

tributed and is as "efficient" as the maximum likelihood estimator based on the full model if the full model is the most parsimonious true model (i.e., if $\beta \neq 0$), and is more "efficient" (namely, as "efficient" as the maximum likelihood estimator based on the restricted model) if the restricted model is the most parsimonious one (i.e., if $\beta = 0$). This seems too good to be true, and, in fact, it is! Although the result in (5) is mathematically correct, it is a delusion to believe that it carries much statistical meaning. Before we explore this in detail, a little reflection shows that the post-model-selection estimator $\tilde{\alpha}$ is nothing else than a variant of Hodges' so-called superefficient estimator (cf. Lehmann and Casella, 1998, pp. 440–443).[10] It is remarkable that estimators such as Hodges' estimator, which was constructed in 1951 as an artificial counterexample to the belief that any asymptotically normally distributed estimator has an asymptotic variance that can not fall below the (asymptotic) Cramér–Rao bound, have nowadays come to some prominence in the guise of post-model-selection estimators based on a consistent model selection procedure (and of other related estimators; see Section 3). It is equally remarkable that some of the lessons learned from Hodges' counterexample seem not to have been received in the model selection literature in the intervening years:[11] The actual finite-sample behavior of $\tilde{\alpha}$ is *not* properly reflected by the (pointwise) asymptotic results; in fact, these results can be highly misleading *regardless* of the sample size and tend to paint an overly optimistic picture of the performance of the estimator. Mathematically speaking, the culprit is nonuniformity (w.r.t. the true parameter vector (α, β)) in the convergence of the finite-sample distributions to the corresponding asymptotic distributions; cf. the warning already issued in Pötscher (1991) in the discussion following Lemma 1 and also in Section 4, Remark (iii), of that paper.

In the simple example discussed here even a finite-sample analysis is possible that allows us to nicely showcase the problems involved.[12] We begin with a closer look at the probability $P_{n,\alpha,\beta}(\hat{M} = M_0)$ of selecting the most parsimonious true model. From (3) this probability equals $\Phi(c) - \Phi(-c)$ if $\beta = 0$, which—in accordance with (4)—goes to unity as sample size increases because we have assumed $c \to \infty$ in this section. In case $\beta \neq 0$, the probability equals $1 - (\Phi(c - \sqrt{n}\beta/\sigma_\beta) - \Phi(-c - \sqrt{n}\beta/\sigma_\beta))$ and—again in accordance with (4)—converges to unity as $n \to \infty$. This is so because $c/\sqrt{n} \to 0$, so that the arguments of the Φ-functions in this formula converge either both to $+\infty$ or both to $-\infty$. Nevertheless, the probability of selecting the most parsimonious true model can be very small for *any* given sample size if $\beta \neq 0$ is close to zero. In that case, we see that this probability is close to $1 - (\Phi(c) - \Phi(-c))$, which in turn is close to zero because of $c \to \infty$. More precisely, if $\beta \neq 0$ equals $\zeta\sigma_\beta c/\sqrt{n}$, $|\zeta| < 1$, then—despite (4)—the probability of selecting the most parsimonious true model in fact converges to zero![13] That is, the consistent model selection procedure is completely "blind" to certain deviations from the restricted model that are of the order $c/\sqrt{n}$. In particular, this reveals that the convergence in (4) is decidedly nonuniform w.r.t. β: In other words, for the

https://doi.org/10.1017/S0266466605050036 Published online by Cambridge University Press

asymptotics to "kick in" in (4) arbitrarily large sample sizes are needed depending on the value of the parameter β. This means that $\hat{M}$, although being consistent for M_0, is not *uniformly* consistent (not even locally). (This is in fact true for *any* consistent model selection procedure; see Remark 4.4.) We illustrate this now numerically. In the following discussion, it proves useful to write γ as shorthand for $(\sqrt{n}/\sigma_\beta)\beta$, i.e., to reparameterize β as $\beta = (\sigma_\beta/\sqrt{n})\gamma$. As a function of γ, the probability of selecting the unrestricted model (which is the most parsimonious true model in case $\beta \neq 0$) is pictured in Figure 1. Recall that with the choice $c = \sqrt{\log n}$ our model selection procedure coincides with the minimum BIC method.

Figure 1 confirms that the probability of selecting the correct model can be very small if $\beta \neq 0$ is of the order $O(1/\sqrt{n})$ and also suggests that this effect even gets stronger as the sample size increases. The latter observation is explained by the fact that the probability of selecting the correct model converges to zero not only for $\beta \neq 0$ of the order $O(1/\sqrt{n})$ but even for $\beta \neq 0$ of larger order, namely, for β of the form $\zeta\sigma_\beta c/\sqrt{n}$, $|\zeta| < 1$; cf. Proposition A.1 in Appendix A. Furthermore, we can also calculate, for given $\beta \neq 0$, how many data points are needed such that the probability of selecting the correct (i.e., the unrestricted) model is at least 0.9, say. With $c = \sqrt{\log n}$ as in Figure 1, we obtain: If $\beta/\sigma_\beta = 1$, then a sample of size $n \geq 8$ is needed; if $\beta/\sigma_\beta = \frac{1}{2}$, one needs $n \geq 42$; if $\beta/\sigma_\beta = \frac{1}{4}$, one needs $n \geq 207$; and if $\beta/\sigma_\beta = \frac{1}{8}$, then $n \geq 977$ is required. This demonstrates that the required sample size heavily depends on the unknown β and increases without bound as β gets closer to zero.

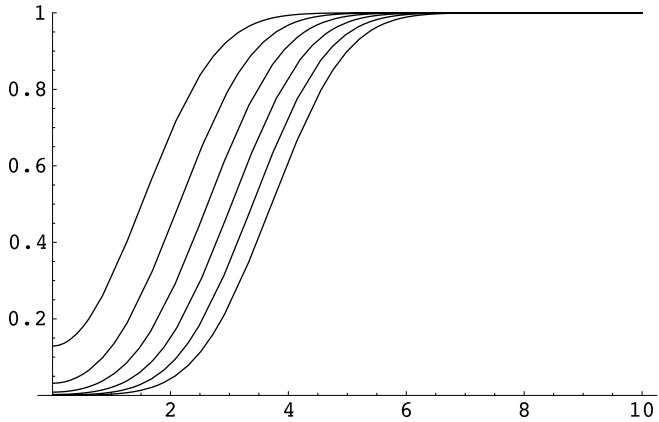

FIGURE 1. Finite-sample model selection probability. The probability of selecting the unrestricted model as a function of $\gamma = \sqrt{n}\beta/\sigma_\beta$ for various values of n, where we have taken $c = \sqrt{\log n}$. Starting from the top, the curves show $P_{n,\alpha,\beta}(\hat{M} = U)$ for $n = 10^k$ for $k = 1,2,\ldots,6$. Note that $P_{n,\alpha,\beta}(\hat{M} = U)$ is independent of α and symmetric around zero in β or, equivalently, γ.

https://doi.org/10.1017/S0266466605050036 Published online by Cambridge University Press

The phenomenon discussed here occurs only if the parameter $\beta \neq 0$ is "small" in absolute value in the sense that it goes to zero of a certain order.[14] It might then be tempting to argue that in such a case erroneously selecting the restricted model is not necessarily detrimental as the restricted model is only "marginally" misspecified: In particular, the estimator $\tilde{\alpha}$ is consistent, even uniformly consistent (cf. Proposition A.9 in Appendix A), and satisfies $\tilde{\alpha} - \alpha = O_P(n^{-1/2})$ as $n \to \infty$ (where O_P is understood relative to $P_{n,\alpha,\beta}$ for fixed α and β). However, given that the consistent model selection procedure is "blind" to deviations from the restricted model of the order $1/\sqrt{n}$ (and even to deviations of larger order), it should not come as a surprise that the phenomenon discussed previously crops up again in the distribution of $\sqrt{n}(\tilde{\alpha} - \alpha)$. Recall that, as a consequence of (5), $\sqrt{n}(\tilde{\alpha} - \alpha)$ is asymptotically normally distributed with mean zero and variance equal to the asymptotic variance of the restricted least-squares estimator if $\beta = 0$ and equal to the asymptotic variance of the unrestricted least-squares estimator if $\beta \neq 0$. However, in finite samples—regardless of how large—we get a completely different picture: From Leeb (2003b), we obtain that the finite-sample density of $\sqrt{n}(\tilde{\alpha} - \alpha)$ is given by

$$\begin{aligned} g_{n,\alpha,\beta}(u) = {} & \sigma_\alpha^{-1}(1-\rho^2)^{-1/2}\phi\big(u(1-\rho^2)^{-1/2}/\sigma_\alpha + \rho(1-\rho^2)^{-1/2}\sqrt{n}\beta/\sigma_\beta\big) \\ & \times \Delta(\sqrt{n}\beta/\sigma_\beta, c) \\ & + \sigma_\alpha^{-1}\left(1 - \Delta\left(\frac{\sqrt{n}\beta/\sigma_\beta + \rho u/\sigma_\alpha}{\sqrt{1-\rho^2}}, \frac{c}{\sqrt{1-\rho^2}}\right)\right)\phi(u/\sigma_\alpha), \end{aligned} \tag{6}$$

where $\phi(\cdot)$ denotes the standard normal probability density function (p.d.f.). Furthermore, we have used $\Delta(a,b)$ as shorthand for $\Phi(a+b) - \Phi(a-b)$, where Φ denotes the standard normal c.d.f. Note that $\Delta(a,b)$ is symmetric in its first argument. The finite-sample density of $\sqrt{n}(\tilde{\alpha} - \alpha)$ does not depend on α and is the sum of two terms: The first term is the density of $\sqrt{n}(\hat{\alpha}(R) - \alpha)$ multiplied by the probability of selecting the restricted model. The second term is a "deformed" version of the density of $\sqrt{n}(\hat{\alpha}(U) - \alpha)$, where the deformation factor is given by the $1 - \Delta(\cdot,\cdot)$-term.[15] Figure 2 gives an example of the possible shapes of the density of $\sqrt{n}(\tilde{\alpha} - \alpha)$.

Two of the densities in Figure 2 are unimodal: The one with the larger mode arises for $\beta/\sigma_\beta = 0$ and is quite close to the (normal) density of $\sqrt{n}(\hat{\alpha}(R) - \alpha)$ corresponding to the restricted model. The reason for this is that the probability $\Delta(0,c)$ of selecting the restricted model is large, namely, 0.968, and hence the first term in (6) is the dominant one. The density with the smaller mode arises for $\beta/\sigma_\beta = 0.5$ and closely resembles the density of $\sqrt{n}(\hat{\alpha}(U) - \alpha)$ corresponding to the unrestricted model. The reason here is (i) that the probability of selecting the unrestricted model is large, namely, 0.998, and hence the second term in (6) is dominant and (ii) that this dominant term is approximately Gaussian; more precisely, the second term in (6) is approximately equal to $\phi(u)(1 - \Delta(7 + 0.98u, 3))$, which differs from $\phi(u)$ in absolute value by less

https://doi.org/10.1017/S0266466605050036 Published online by Cambridge University Press

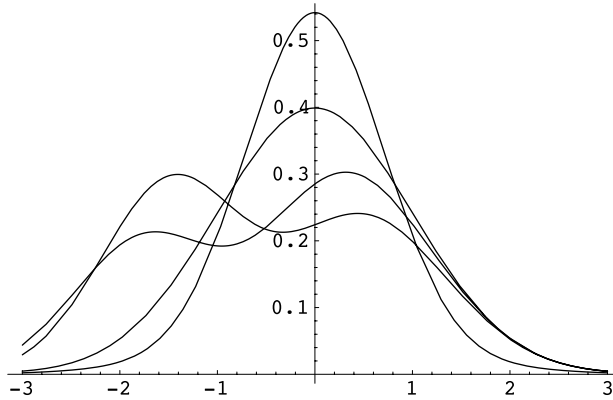

FIGURE 2. Finite-sample densities. The density $g_{n,\alpha,\beta}$ of $\sqrt{n}(\tilde{\alpha} - \alpha)$ for various values of β/σ_β. For the graphs, we have taken $n = 100$, $c = \sqrt{\log n}$, $\rho = 0.7$, and $\sigma_\alpha^2 = 1$. The four curves correspond to β/σ_β equal to 0, 0.21, 0.25, and 0.5 and are discussed in the text.

than 0.002. The bimodal densities correspond to the cases $\beta/\sigma_\beta = 0.21$ and $\beta/\sigma_\beta = 0.25$. In both cases, the left-hand mode reflects the contribution of the first term in (6) whereas the right-hand mode reflects the contribution of the second term. The height of the left-hand mode is proportional to the probability of selecting the restricted model, which is larger for $\beta/\sigma_\beta = 0.21$ than for $\beta/\sigma_\beta = 0.25$. In summary, we see that the finite-sample distribution of $\sqrt{n}(\tilde{\alpha} - \alpha)$ depends heavily on the value of the unknown parameter β (through β/σ_β) and that it is far from its Gaussian large-sample limit distribution for certain values of β. The same phenomenon is also found if we repeat the calculations for other sample sizes n, regardless of how large n is. In other words: Although the distribution of $\sqrt{n}(\tilde{\alpha} - \alpha)$ is approximately Gaussian for each *given* (α, β) and sufficiently large sample size, the amount of data required to achieve a *given* accuracy of approximation depends on the unknown β. In the example presented in Figure 2, a sample size of 100 appears to be sufficient for the normal approximation predicted by pointwise asymptotic theory to be reasonably accurate in the cases $\beta/\sigma_\beta = 0$ and $\beta/\sigma_\beta = 0.5$, whereas it is clearly insufficient in case $\beta/\sigma_\beta = 0.21$ or $\beta/\sigma_\beta = 0.25$.

How can this be reconciled with the result mentioned earlier that $\sqrt{n}(\tilde{\alpha} - \alpha)$ has an asymptotic normal distribution with mean zero and appropriate variance? The crucial observation again is that this limit result is a pointwise one; i.e., it holds for each *fixed* value of the parameter vector (α, β) individually but does not hold uniformly w.r.t. (α, β) (in fact, not even locally uniformly): While it is easy to see that for every $u \in \mathbb{R}$ the density $g_{n,\alpha,\beta}(u)$ given by (6) converges to the appropriate normal density for each *fixed* (α, β), it is equally easy to see (cf. Proposition A.2 in Appendix A) that (6) has a different asymptotic

https://doi.org/10.1017/S0266466605050036 Published online by Cambridge University Press

behavior if, e.g., $\beta = \sigma_\beta \gamma/\sqrt{n}$ with $\gamma \neq 0$. In this case (6) converges to a *shifted* version of the density of the asymptotic distribution of $\sqrt{n}(\hat{\alpha}(R) - \alpha)$, the shift being controlled by γ. Yet another asymptotic behavior is obtained if we consider $\beta = \sigma_\beta \gamma_n/\sqrt{n}$ with $\gamma_n \to \infty$ (or $\gamma_n \to -\infty$) but $\gamma_n = o(c)$. Then $g_{n,\alpha,\beta}(u)$ even converges to zero for every $u \in \mathbb{R}$! That is, the distribution of $\sqrt{n}(\tilde{\alpha} - \alpha)$ does not "stabilize" as sample size increases but—loosely speaking—"escapes" to ∞ or $-\infty$ (depending on the sign of γ_n); in fact, $\sqrt{n}(\tilde{\alpha} - \alpha) \to \infty$ or $-\infty$ in $P_{n,\alpha,\beta}$-probability. More complicated asymptotic behavior is in fact possible and is described in Proposition A.2 in Appendix A.[16] (To simplify matters the rather special case $\rho_\infty = 0$ is excluded from the preceding discussion; cf. Remark 4.6 for some comments on this case. However, note that Proposition A.2 also covers the case $\rho_\infty = 0$.)

We are now in a position to analyze the actual coverage properties of confidence intervals that are constructed "as usual," thereby ignoring the presence of model selection (this step seemingly being justified by a reference to (5)). Let $\mathcal{I}$ denote the "naive" confidence interval that is given by the usual confidence interval in the restricted (unrestricted) model if the restricted (unrestricted) model is selected. That is,

$$\mathcal{I} = [\tilde{\alpha} - z_\eta n^{-1/2} \sigma_\alpha (1 - \rho^2)^{1/2}, \tilde{\alpha} + z_\eta n^{-1/2} \sigma_\alpha (1 - \rho^2)^{1/2}] \tag{7}$$

if $\hat{M} = R$ and

$$\mathcal{I} = [\tilde{\alpha} - z_\eta n^{-1/2} \sigma_\alpha, \tilde{\alpha} + z_\eta n^{-1/2} \sigma_\alpha] \tag{8}$$

if $\hat{M} = U$ where $1 - \eta$ denotes the *nominal* coverage probability and z_η is the $(1 - \eta/2)$ quantile of a standard normal distribution. In view of (2), the *actual* coverage probability satisfies

$$\begin{aligned} P_{n,\alpha,\beta}(\alpha \in \mathcal{I}) &= P_{n,\alpha,\beta}(\alpha \in \mathcal{I} | \hat{M} = R) P_{n,\alpha,\beta}(\hat{M} = R) \\ &\quad + P_{n,\alpha,\beta}(\alpha \in \mathcal{I} | \hat{M} = U) P_{n,\alpha,\beta}(\hat{M} = U). \end{aligned} \tag{9}$$

Using the remark in note 15 in the notes section, it is an elementary calculation to obtain

$$\begin{aligned} &P_{n,\alpha,\beta}(\alpha \in \mathcal{I}) \\ &\quad = \Delta(\rho(1 - \rho^2)^{-1/2} \sqrt{n}\beta/\sigma_\beta, z_\eta) \Delta(\sqrt{n}\beta/\sigma_\beta, c) \\ &\qquad + \int_{-z_\eta}^{z_\eta} (1 - \Delta((\sqrt{n}\beta/\sigma_\beta + \rho u)(1 - \rho^2)^{-1/2}, c(1 - \rho^2)^{-1/2})) \phi(u)\, du. \end{aligned} \tag{10}$$

Note that the coverage probability does not depend on α and is symmetric around zero as a function of β. Because of (5) and the attending discussion, pointwise asymptotic theory tells us that the coverage probability $P_{n,\alpha,\beta}(\alpha \in \mathcal{I})$ con-

https://doi.org/10.1017/S0266466605050036 Published online by Cambridge University Press

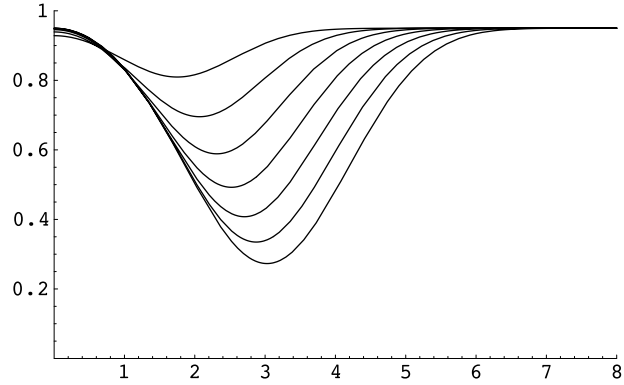

FIGURE 3. Finite-sample coverage probabilities. The coverage probability of the "naive" confidence interval $\mathcal{I}$ with nominal confidence level $1-\eta = 0.95$ as a function of $\gamma = \sqrt{n}\beta/\sigma_\beta$ for various values of n, where we have taken $c = \sqrt{\log n}$ and $\rho = 0.7$. The curves are given for $n = 10^k$ for $k = 1,2,\ldots,7$; larger sample sizes correspond to curves with a smaller minimal coverage probability.

verges to $1-\eta$ for every (α,β). However, the plots of the coverage probability given in Figure 3 speak another language.

We see that the *actual* coverage probability of the "naive" interval $\mathcal{I}$ is often far below its nominal level of 0.95, sometimes falling below 0.3. Figure 3 also suggests that this phenomenon gets more pronounced when sample size increases! In fact, it is not difficult to see that the minimal coverage probability of $\mathcal{I}$ converges to zero as sample size increases and not to the nominal coverage probability $1-\eta$ as one might have hoped for (except possibly in the relatively special case $\rho_\infty = 0$); cf. also Kabaila (1995). To see this, note that

$$\min_{\alpha,\beta} P_{n,\alpha,\beta}(\alpha \in \mathcal{I}) \le P_{n,\alpha,\sigma_\beta\gamma_n/\sqrt{n}}(\alpha \in \mathcal{I}),$$

where α is arbitrary and γ_n is chosen such that $\gamma_n \to \infty$ (or $\gamma_n \to -\infty$) and $\gamma_n = o(c)$. (The r.h.s. in the preceding inequality does actually not depend on α in view of (10).) Because $P_{n,\alpha,\sigma_\beta\gamma_n/\sqrt{n}}(\hat{M} = U)$ converges to zero as discussed earlier (cf. Proposition A.1 in Appendix A), we arrive—using (9) and (10)—at

$$\begin{aligned}
&\lim_{n\to\infty} \min_{\alpha,\beta} P_{n,\alpha,\beta}(\alpha \in \mathcal{I}) \\
&\quad \le \lim_{n\to\infty} P_{n,\alpha,\sigma_\beta\gamma_n/\sqrt{n}}(\alpha \in \mathcal{I} | \hat{M} = R) P_{n,\alpha,\sigma_\beta\gamma_n/\sqrt{n}}(\hat{M} = R) \\
&\quad = \lim_{n\to\infty} \Delta(\rho(1-\rho^2)^{-1/2}\gamma_n, z_\eta)\Delta(\gamma_n, c) = 0,
\end{aligned}$$

the last equality being true because $|\gamma_n| \to \infty$ (and because we have excluded the case $\rho_\infty = 0$).

https://doi.org/10.1017/S0266466605050036 Published online by Cambridge University Press

We finally illustrate the impact of model selection on the (scaled) bias and the (scaled) mean-squared error of the estimator (again excluding for simplicity of discussion the case $\rho_\infty = 0$). Let *Bias* denote the expectation and *MSE* the second moment of $\sqrt{n}(\tilde{\alpha} - \alpha)$. We discuss the bias first. An explicit formula for the bias can be obtained from (6) by a tedious but straightforward computation and is given by

$$Bias = -\rho\sigma_\alpha[(\sqrt{n}\beta/\sigma_\beta)\Delta(\sqrt{n}\beta/\sigma_\beta, c) - \phi(\sqrt{n}\beta/\sigma_\beta - c) + \phi(\sqrt{n}\beta/\sigma_\beta + c)]. \quad \textbf{(11)}$$

A pointwise (i.e., for fixed (α, β)) asymptotic analysis tells us that this bias vanishes asymptotically.[17] In Figure 4 we have computed this bias numerically as a function of $\gamma = \sqrt{n}\beta/\sigma_\beta$. Note that the bias is independent of α and antisymmetric around zero in β or, equivalently, γ (and hence is shown only for $\gamma \geq 0$).

Figure 4 demonstrates that—contrary to the prediction of pointwise asymptotic theory—the bias can be quite substantial if β is of the order $O(1/\sqrt{n})$ and that this effect gets more pronounced as the sample size increases (the reason for this discrepancy again being nonuniformity in the pointwise asymptotic results). An asymptotic analysis of (11) using $\beta = \sigma_\beta\gamma/\sqrt{n}$ with $\gamma \neq 0$ shows that the bias converges to $-\sigma_\alpha\rho_\infty\gamma$ (see Proposition A.4 in Appendix A for more information). Note that this limit corresponds to the "envelope" of the finite-sample bias curves (for all n) as indicated in Figure 4. Furthermore, if $\beta = \sigma_\beta\gamma_n/\sqrt{n}$ with $\gamma_n \to \infty$ (or $\gamma_n \to -\infty$) but $\gamma_n = o(c)$, the asymptotic analysis in Proposition A.4 even shows that the bias converges to $\pm\infty$, the sign

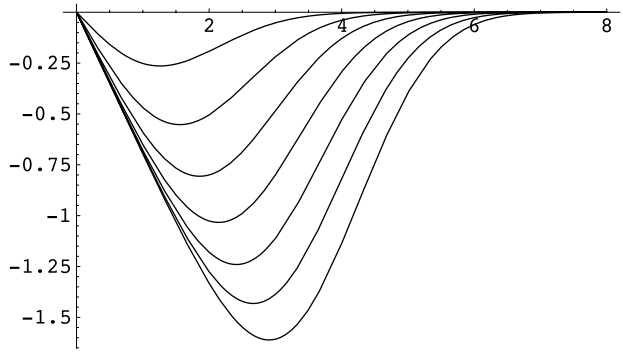

FIGURE 4. Finite-sample bias. The expectation of $\sqrt{n}(\tilde{\alpha} - \alpha)$, i.e., the (scaled) bias of the post-model-selection estimator for α, as a function of $\gamma = \sqrt{n}\beta/\sigma_\beta$ for various values of n, where we have taken $c = \sqrt{\log n}$, $\rho = 0.7$, and $\sigma_\alpha^2 = 1$. The curves are given for $n = 10^k$ for $k = 1,2,\ldots,7$; larger sample sizes correspond to curves with larger maximal absolute biases.

https://doi.org/10.1017/S0266466605050036 Published online by Cambridge University Press

depending on the sign of γ_n. As a consequence, the maximal absolute bias in fact grows without bound as sample size increases!

Turning to the *MSE* we encounter a similar situation. Using the fact that the test statistic $|\sqrt{n}\hat{\beta}(U)/\sigma_\beta|$ is independent of $\hat{\alpha}(R)$ (e.g., Leeb and Pötscher, 2003a, Proposition 3.1) and that $\hat{\alpha}(R) = \hat{\alpha}(U) - \rho(\sigma_\alpha/\sigma_\beta)\hat{\beta}(U)$, the *MSE* can be computed explicitly to be

$$MSE = \sigma_\alpha^2 + \sigma_\alpha^2 \rho^2 \left[\begin{array}{l} (c + \sqrt{n}\beta/\sigma_\beta)\phi(c + \sqrt{n}\beta/\sigma_\beta) \\ \quad + (c - \sqrt{n}\beta/\sigma_\beta)\phi(c - \sqrt{n}\beta/\sigma_\beta) \\ \quad + ((n\beta^2/\sigma_\beta^2) - 1)(\Phi(c - \sqrt{n}\beta/\sigma_\beta) \\ \qquad\qquad - \Phi(-c - \sqrt{n}\beta/\sigma_\beta)) \end{array} \right]. \tag{12}$$

Alternatively, the preceding formula can also be obtained by brute force integration from the density (6) or from Theorems 2.2 and 4.1 in Magnus (1999). The *MSE* is independent of α. A pointwise asymptotic analysis tells us that *MSE* converges to the asymptotic variance $\sigma_{\alpha,\infty}^2(1 - \rho_\infty^2)$ of $\sqrt{n}(\hat{\alpha}(R) - \alpha)$ if $\beta = 0$ and to the asymptotic variance $\sigma_{\alpha,\infty}^2$ of $\sqrt{n}(\hat{\alpha}(U) - \alpha)$ if $\beta \neq 0$.[18] Again, however, the finite-sample mean-squared error exhibits a totally different behavior, regardless how large sample size is (as a result of nonuniformity in the pointwise asymptotics). This can be gleaned from Figure 5: The maximal mean-squared error is much larger than the mean-squared error of the unrestricted least-squares estimator that is constant and equal to $\sigma_\alpha^2 = 1$. As Figure 5 suggests, the maximal mean-squared error diverges to infinity as sam-

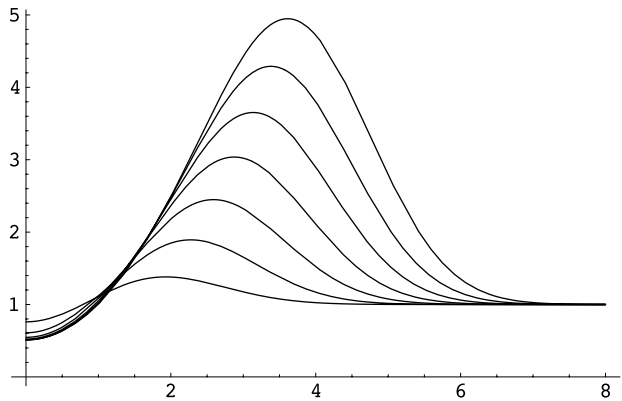

FIGURE 5. Finite-sample mean-squared error. The second moment of $\sqrt{n}(\tilde{\alpha} - \alpha)$, i.e., the (scaled) mean-squared error of the post-model-selection estimator for α, as a function of $\gamma = \sqrt{n}\beta/\sigma_\beta$ for various values of n, where we have taken $c = \sqrt{\log n}$, $\rho = 0.7$, and $\sigma_\alpha^2 = 1$. The curves are given for $n = 10^k$ for $k = 1,2,\ldots,7$; larger sample sizes correspond to curves with larger maximal mean-squared error.

https://doi.org/10.1017/S0266466605050036 Published online by Cambridge University Press

ple size increases, whereas the mean-squared error of $\sqrt{n}(\hat{\alpha}(U) - \alpha)$ stays bounded (it converges to $\sigma^2_{\alpha,\infty}$). This is well known for the Hodges estimator (e.g., Lehmann and Casella, 1998, p. 442). For the mean-squared error of $\sqrt{n}(\tilde{\alpha} - \alpha)$ this follows of course immediately from the fact noted previously that the bias diverges to $\pm\infty$ when setting $\beta = \sigma_\beta \gamma_n / \sqrt{n}$ with $\gamma_n \to \infty$ (or $\gamma_n \to -\infty$) but $\gamma_n = o(c)$. (The phenomenon that the maximal absolute bias and hence the maximal mean-squared error diverge to infinity holds for post-model-selection estimators based on consistent model selection procedures in general; see Remark 4.1, Appendix C; and Yang (2003).)

2.2. The Conservative Model Selection Framework

Generally speaking, post-model-selection estimators based on conservative model selection procedures are subject to phenomena similar to the ones observed in Section 2.1 for post-model-selection estimators based on consistent procedures. In particular, the finite-sample behavior of both types of post-model-selection estimators is governed by exactly the same formulas, because the finite-sample behavior is clearly not much impressed by what we fancy about the behavior of the model selection procedure at fictitious sample sizes other than n (e.g., what we fancy about the behavior of the cutoff point c as a function of n). Cf. the discussion immediately preceding Section 2.1. Not surprisingly, some differences arise in the asymptotic theory.

In this section we consider the same model selection procedure and post-model-selection estimator $\tilde{\alpha}$ as before, except that we now assume the cutoff point c to be independent of sample size n.[19] This results in a conservative model selection procedure (that is not consistent).[20] As just noted, the finite-sample distribution, the expectation, and the second moment of $\sqrt{n}(\tilde{\alpha} - \alpha)$ are again given by (6), (11), and (12), respectively. Also, the model selection probabilities and the coverage probability of the "naive" confidence interval are given by the same formulas as before. As a consequence, all conclusions drawn from the finite-sample formulas in Section 2.1 remain valid here: The finite-sample distribution of the post-model-selection estimator is often decidedly nonnormal, and the standard asymptotic approximations derived on the presumption of an a priori given model are inappropriate. In particular, the actual coverage probability of the "naive" confidence interval is often much smaller than the nominal coverage probability. Finally, the bias can be substantial, and the mean-squared error can by far exceed the mean-squared error of the unrestricted estimator.

We briefly discuss the asymptotic behavior next.[21] A much more detailed treatment covering more general model selection procedures and more general models can be found in Pötscher (1991), Leeb and Pötscher (2003a), and Leeb (2003a, b). The pointwise limiting behavior of the model selection probabilities can be easily read off from the finite-sample formula (3): $\lim_{n\to\infty} P_{n,\alpha,\beta}(\hat{M} = R) = 0$ if $\beta \neq 0$ and $\lim_{n\to\infty} P_{n,\alpha,\beta}(\hat{M} = R) = \Phi(c) - \Phi(-c) < 1$ if $\beta = 0$, reflecting

https://doi.org/10.1017/S0266466605050036 Published online by Cambridge University Press

the fact that the model selection procedure is conservative but not consistent. As in the case of consistent model selection procedures, this convergence is not uniform w.r.t. β. In contrast to consistent model selection procedures (cf. Proposition A.1 in Appendix A), the behavior under sample-size-dependent parameters (α_n, β_n) is quite simple: If $\sqrt{n}\beta_n/\sigma_\beta \to \gamma$, $|\gamma| < \infty$, then $\lim_{n\to\infty} P_{n,\alpha_n,\beta_n}(\hat{M} = R) = \Phi(c - \gamma) - \Phi(-c - \gamma)$. (If $\sqrt{n}|\beta_n|/\sigma_\beta \to \infty$, then the limit is zero; i.e., the asymptotic behavior is identical to the asymptotic behavior under fixed $\beta \neq 0$.) In particular, the asymptotic analysis confirms what we already know from the finite-sample analysis, namely, that the probability of erroneously selecting the restricted model can be substantial, namely, if $|\gamma|$ is small. However, in contrast to consistent model selection procedures, this probability does not converge to unity as sample size increases. It is also interesting to note that deviations from the restricted model such as $\beta = \zeta\sigma_\beta c_n/\sqrt{n}$ with $|\zeta| < 1$ and $c_n \to \infty$, $c_n/\sqrt{n} \to 0$, that can not be detected by a consistent model selection procedure using cutoff point c_n (cf. Proposition A.1 and note 14 in the notes section) can be detected with probability approaching unity by a conservative procedure using a fixed cutoff point c. Consequently and not surprisingly, conservative model selection procedures are more powerful than consistent model selection procedures in the sense that they are less likely to erroneously select an incorrect model for large sample sizes. (Needless to say this advantage of the conservative procedure is paid for by a larger probability of selecting an overparameterized model.)

Turning to the post-model-selection estimator $\tilde{\alpha}$ itself, it is obvious that now conditions (4) and (5) are no longer satisfied;[22] as a consequence, and in contrast to the case of consistent model selection procedures, the pointwise asymptotic distribution now captures some of the effects of model selection and no longer coincides with the usual asymptotic distribution that applies in the absence of model selection. This can easily be seen from (2): Whereas in the case of consistent model selection procedures, regardless of the value of β, only one of the two terms in (2) survives asymptotically and the corresponding conditioning event becomes a set of probability one asymptotically and hence has no effect, for conservative procedures both terms do not vanish in the limit if $\beta = 0$. Hence, the pointwise asymptotic limit captures some of the effects of the model selection step, at least in the case when the restricted model is correct. (In that sense the asymptotic framework that views a given model selection procedure as embedded in a sequence of conservative procedures has some advantage over the framework considered in Section 2.1.) More precisely, the pointwise asymptotic distribution of $\sqrt{n}(\tilde{\alpha} - \alpha)$ has a density given by $\sigma_{\alpha,\infty}^{-1}\phi(u/\sigma_{\alpha,\infty})$ if $\beta \neq 0$ and given by

$$\sigma_{\alpha,\infty}^{-1}(1-\rho_\infty^2)^{-1/2}\phi(u(1-\rho_\infty^2)^{-1/2}/\sigma_{\alpha,\infty})\Delta(0,c) + \sigma_{\alpha,\infty}^{-1}\left[1 - \Delta\left(\frac{\rho_\infty u/\sigma_{\alpha,\infty}}{\sqrt{1-\rho_\infty^2}}, \frac{c}{\sqrt{1-\rho_\infty^2}}\right)\right]\phi(u/\sigma_{\alpha,\infty}) \quad \textbf{(13)}$$

https://doi.org/10.1017/S0266466605050036 Published online by Cambridge University Press

if $\beta = 0$. Note that (13) bears some resemblance to the finite-sample distribution (6). However, the pointwise asymptotic distribution does not capture all the effects present in the finite-sample distribution, especially if $\beta \neq 0$; in particular, the convergence is not uniform w.r.t. β (except in trivial cases such as $\rho_\infty = 0$); cf. Corollary 5.5 in Leeb and Pötscher (2003a), Remark 6.6 in Leeb and Pötscher (2003b), and note 16. A much better approximation, capturing all the essential features of the finite-sample distribution, is obtained by the asymptotic distribution under sample-size dependent parameters (α_n, β_n) with $\sqrt{n}\beta_n/\sigma_\beta \to \gamma$, $|\gamma| < \infty$: This asymptotic distribution has a density of the form

$$\sigma_{\alpha,\infty}^{-1}(1-\rho_\infty^2)^{-1/2}\phi(u(1-\rho_\infty^2)^{-1/2}/\sigma_{\alpha,\infty} + \rho_\infty(1-\rho_\infty^2)^{-1/2}\gamma)\Delta(\gamma, c)$$
$$+ \sigma_{\alpha,\infty}^{-1}\left[1 - \Delta\left(\frac{\gamma + \rho_\infty u/\sigma_{\alpha,\infty}}{\sqrt{1-\rho_\infty^2}}, \frac{c}{\sqrt{1-\rho_\infty^2}}\right)\right]\phi(u/\sigma_\alpha). \qquad \textbf{(14)}$$

This follows either as a special case of Proposition 5.1 of Leeb (2003b) (cf. also Leeb and Pötscher, 2003a, Proposition 5.3 and Corollary 5.4) or can be gleaned directly from (6). (If $\sqrt{n}|\beta_n|/\sigma_\beta \to \infty$, then the limit has the form $\sigma_{\alpha,\infty}^{-1}\phi(u/\sigma_{\alpha,\infty})$.)[23] Observe that (14) follows the same formula as the finite-sample density (6), except that σ_α and ρ have been replaced by their respective limits $\sigma_{\alpha,\infty}$ and ρ_∞ and that $\sqrt{n}\beta/\sigma_\beta$ has been replaced by γ.

Consider next the asymptotic behavior of the actual coverage probability of the "naive" confidence interval $\mathcal{I}$ given by (7) and (8). The pointwise limit of the actual coverage probability has been studied in Pötscher (1991, Sect. 3.3). In contrast to the case of consistent model selection procedures, it turns out to be less than the nominal coverage probability in case the restricted model is correct. However, this pointwise asymptotic result, although hinting at the problem, still gives a much too optimistic picture when compared with the actual finite-sample coverage probability. The large-sample minimal coverage probability of the "naive" confidence interval has been studied in Kabaila and Leeb (2004). Although it does not equal zero as in the case of consistent model selection procedures, it turns out to be often much smaller than the nominal coverage probability $1 - \eta$ (as in Figure 3); see Kabaila and Leeb (2004) for more details.

We finally turn to the bias and mean-squared error of $\sqrt{n}\tilde{\alpha}$. Under the sequence of parameters (α_n, β_n) with $\sqrt{n}\beta_n/\sigma_\beta \to \gamma$, $|\gamma| < \infty$, it is readily seen from (11) that the bias converges to

$$-\rho_\infty\sigma_{\alpha,\infty}[\gamma\Delta(\gamma, c) - \phi(\gamma - c) + \phi(\gamma + c)].$$

The pointwise asymptotics corresponds to the cases $\gamma = 0$ and $\gamma = \pm\infty$ (with the convention that $\pm\infty\Delta(\pm\infty, c) = 0$ and $\phi(\pm\infty) = 0$) and results in a zero limiting bias. However, the maximal bias can be quite substantial if β is of the order $O(1/\sqrt{n})$. In contrast to the case of consistent model selection proce-

https://doi.org/10.1017/S0266466605050036 Published online by Cambridge University Press

dures, the maximal bias does not go to infinity (in absolute value) as $n \to \infty$ but remains bounded. (It is perhaps somewhat ironic—although not surprising—that consistent model selection procedures that look perfect in a pointwise asymptotic analysis lead in fact to more heavily distorted post-model-selection estimators than conservative model selection procedures.) The limiting mean-squared error under (α_n, β_n) as before is easily seen to be given by

$$\sigma^2_{\alpha,\infty} + \sigma^2_{\alpha,\infty}\rho^2_\infty \left[\begin{array}{l} (c+\gamma)\phi(c+\gamma) + (c-\gamma)\phi(c-\gamma) \\ \quad + (\gamma^2 - 1)(\Phi(c-\gamma) - \Phi(-c-\gamma)) \end{array} \right],$$

the pointwise asymptotics again corresponding to the cases $\gamma = 0$ and $\gamma = \pm\infty$ (with the convention that $\infty\Delta(\pm\infty, c) = 0$ and $\pm\infty\phi(\pm\infty) = 0$). In contrast to the case of consistent model selection procedures, the pointwise limit of *MSE* captures some (but not all) of the effects of model selection and hence no longer coincides with the asymptotic variance of the infeasible "estimator" $\hat{\alpha}(M_0)$. Also, in contrast to the case of consistent model selection procedures, the maximal mean-squared error does not go off to infinity as $n \to \infty$, but rather it remains bounded; cf. also Remark 4.1.

2.3. Can One Estimate the Distribution of Post-Model-Selection Estimators?

It transpires from the preceding discussion that the finite-sample distributions (and also the asymptotic distributions) of post-model-selection estimators depend on unknown parameters (i.e., β in the example discussed in this paper), often in a complicated fashion. For inference purposes, e.g., for the construction of confidence sets, estimators for these distributions would be desirable. Consistent estimators for these distributions can typically be constructed quite easily, e.g., by suitably replacing unknown parameters in the large-sample limit distributions by estimators: In the case of the consistent model selection procedure discussed in Section 2.1 a consistent estimator for the finite-sample distribution of $\sqrt{n}(\tilde{\alpha} - \alpha)$ is simply given by the normal distribution $N(0, \sigma^2_\alpha(1 - \rho^2))$, i.e., by the distribution of $\sqrt{n}(\alpha(R) - \alpha)$, if $\hat{M} = R$, and by $N(0, \sigma^2_\alpha)$, i.e., by the distribution of $\sqrt{n}(\alpha(U) - \alpha)$, if $\hat{M} = U$. However, recall from Section 2.1 that the finite-sample distribution of the post-model-selection estimator is not uniformly close to its pointwise asymptotic limit. Hence the suggested estimator (being identical with the pointwise asymptotic distribution except for replacing $\sigma^2_{\alpha,\infty}$ and ρ^2_∞ by σ^2_α and ρ^2) will—although being consistent—not be close to the finite-sample distribution uniformly in the unknown parameters, thus providing a rather useless estimator. In the case of conservative model selection procedures consistent estimators for the finite-sample distribution of the post-model-selection estimator can also be constructed from the pointwise asymptotic distribution by suitably plugging in estimators for unknown quantities; see Leeb and Pötscher (2003b, 2004). How-

https://doi.org/10.1017/S0266466605050036 Published online by Cambridge University Press

ever, again these estimators will be quite useless for the same reason: As discussed in Section 2.2, the convergence of the finite-sample distributions to their (pointwise) large-sample limits is typically not uniform with respect to the underlying parameters, and there is no reason to believe that this nonuniformity will disappear when unknown parameter values in the large-sample limit are replaced by estimators.

A natural reaction to the preceding discussion could be to try the bootstrap or some related resampling procedure such as, e.g., subsampling. Consider first the case of a consistent model selection procedure. Then, in view of (4) and (5), the bootstrap that resamples from the residuals of the selected model certainly provides a consistent estimator for the finite-sample distribution of the post-model-selection estimator. Note that the consistent estimator described in the preceding paragraph can be viewed as a (parametric) bootstrap. The discussion in the previous paragraph then, however, suggests that such estimators based on the bootstrap (or on other resampling procedures such as subsampling), despite being consistent, will be plagued by the nonuniformity issues discussed earlier. Next consider the case where the model selection procedure is conservative (but not consistent). Then the bootstrap will typically not even provide consistent estimators for the finite-sample distribution of the post-model-selection estimator, as the bootstrap can be shown to stay random in the limit (Kulperger and Ahmed, 1992; Knight, 1999, Example 3):[24] Basically the only way one can coerce the bootstrap into delivering a consistent estimator is to resample from a model that has been selected by an *auxiliary* consistent model selection procedure. (The construction of consistent estimators in Leeb and Pötscher, 2003b, 2004, alluded to previously basically follows this route.) In contrast, subsampling will typically deliver consistent estimators. However, the discussion in the preceding paragraph strongly suggests that any such estimator will again suffer from the nonuniformity defect.

A natural question then is how estimators (not necessarily derived from the asymptotic distributions or from resampling considerations) can be found that do not suffer from the nonuniformity defect. In other words, we are asking for estimators $\hat{G}_{n,\alpha,\beta}$ of the finite-sample c.d.f. $G_{n,\alpha,\beta}$ of $\sqrt{n}(\tilde{\alpha} - \alpha)$ that are uniformly consistent, i.e., that satisfy for every $t \in \mathbb{R}$ and every $\delta > 0$

$$\sup_{\alpha,\beta} P_{n,\alpha,\beta}(|\hat{G}_{n,\alpha,\beta}(t) - G_{n,\alpha,\beta}(t)| > \delta) \xrightarrow{n\to\infty} 0.$$

However, it turns out that *no* estimator $\hat{G}_{n,\alpha,\beta}$ can satisfy this requirement (except possibly in the trivial case where $\rho_\infty = 0$). For conservative model selection procedures this is proved in Leeb and Pötscher (2003a, 2004) in a more general framework, including model selection by AIC from a quite arbitrary collection of linear regression models. For a consistent model selection procedure such a result is given in Leeb and Pötscher (2002, Sect. 2.3). In fact, these papers show that the situation is even more dramatic: For every consistent estimator $\hat{G}_{n,\alpha,\beta}$ of $G_{n,\alpha,\beta}$ even

https://doi.org/10.1017/S0266466605050036 Published online by Cambridge University Press

$$\sup_{\alpha,\beta} P_{n,\alpha,\beta}(|\hat{G}_{n,\alpha,\beta}(t) - G_{n,\alpha,\beta}(t)| > \delta) \xrightarrow{n\to\infty} 1$$

holds for suitable $\delta > 0$, and this result is even local in the sense that it holds also if the supremum in the preceding display extends only over suitable balls that shrink at rate $1/\sqrt{n}$.[25] (These "impossibility" results hold for randomized estimators of $G_{n,\alpha,\beta}$ also.)

The preceding "impossibility" results establish in particular that any proposal to estimate the distribution of post-model-selection estimators by whatever resampling procedure (bootstrap, subsampling, etc.) is doomed as any such estimator is necessarily plagued by the nonuniformity defect (if it is consistent at all). On a more general level, an implication of the preceding results is that assessing the variability of post-model-selection estimators (e.g., the construction of valid confidence intervals post model selection) is a harder problem than perhaps expected.[26]

3. RELATED PROCEDURES: SHRINKAGE-TYPE ESTIMATORS AND PENALIZED LEAST-SQUARES

Post-model-selection estimators can be viewed as a discontinuous form of shrinkage estimators. In this section we briefly discuss the relationship between post-model-selection estimators and shrinkage-type estimators and look at the distributional properties of such estimators. Although estimators such as the James–Stein estimator or ridge estimators have a long tradition in econometrics and statistics, a number of shrinkage-type estimators such as the Lasso estimator, the Bridge estimator, and the SCAD estimator are of more recent vintage. In the context of a linear regression model $Y = X\theta + \varepsilon$ many of these estimators can be cast in the form of a penalized least-squares estimator: Let $\hat{\theta}$ be the estimator that is obtained by minimizing the penalized least-squares criterion

$$\sum_{t=1}^{n} (y_t - x_{t.}\theta)^2 + \lambda_n \sum_{j=1}^{k} |\theta_j|^q, \tag{15}$$

where $x_{t.}$ denotes the tth row and k the number of columns of X. This is the class of Bridge estimators introduced by Frank and Friedman (1993), the case $q = 2$ corresponding to the ridge estimator. The member of this class obtained by setting $q = 1$ has been referred to as a Lasso-type estimator by Knight and Fu (2000), because it is closely related to the Lasso of Tibshirani (1996). Knight and Fu (2000) also note that in the context of wavelet regression minimizing (15) with $q = 1$ is known as "basis pursuit," cf. Chen, Donoho, and Saunders (1998). In fact, in the case of diagonal $X'X$ the Lasso-type estimator reduces to soft-thresholding of the coordinates of the least-squares estimator. (We note that in this case hard-thresholding, which obviously is a model selection procedure, can also be represented as a penalized least-squares estimator.)

https://doi.org/10.1017/S0266466605050036 Published online by Cambridge University Press

The SCAD estimator introduced by Fan and Li (2001) is also a penalized least-squares estimator but uses a different penalty term. It is given as the minimizer of

$$\sum_{t=1}^{n} (y_t - x_{t.}\theta)^2 + \sum_{j=1}^{k} p_{\lambda_n}(\theta_j)$$

with a specific choice of p_{λ_n} that we do not reproduce here.

The asymptotic distributional properties of Bridge estimators have been studied in Knight and Fu (2000). Under appropriate conditions on q and on the regularization parameter λ_n, the asymptotic distribution shows features similar to the asymptotic distribution of post-model-selection estimators based on a conservative model selection procedure (e.g., bimodality). Under other conditions on q and λ_n, the Bridge estimator acts more like a post-model-selection estimator based on a consistent procedure. In particular, such a Bridge estimator will estimate zero components of the true θ exactly as zero with probability approaching unity. It hence satisfies an "oracle" property. This is also true for the SCAD estimator of Fan and Li (2001). In view of the discussion in Section 2.1 and the lessons learned from Hodges' estimator, one should, however, not read too much into this property as it can give a highly misleading impression of the properties of these estimators in finite samples.[27]

Another similarity with post-model-selection estimators is the fact that the distribution function or the risk of shrinkage-type estimators often can not be estimated uniformly consistently. See Leeb and Pötscher (2002) for more on this subject.

4. REMARKS

Remark 4.1. In this remark we collect some decision-theoretic facts about post-model-selection estimators. These results could be taken as a starting point for a discussion of whether or not model selection (from submodels of an overall model of fixed finite dimension) can be justified from a decision-theoretic point of view.

1. Sometimes model selection is motivated by arguing that allowing for the selection of models more parsimonious than the overall model would lead to a gain in the precision of the estimate. However, this argument does not hold up to closer scrunity. For example, it is well known in the standard linear regression model $Y = X\theta + \varepsilon$ that the mean-squared error of any given pretest estimator for θ exceeds the mean-squared error of the least-squares estimator $(X'X)^{-1}X'Y$ on parts of the parameter space (Judge and Bock, 1978; Judge and Yancey, 1986; Magnus, 1999). Hence, pretesting does not lead to a global gain (i.e., a gain that holds over the entire parameter space) in mean-squared error over the least-squares estimator

https://doi.org/10.1017/S0266466605050036 Published online by Cambridge University Press

obtained from the overall model. Cf. also the discussion of the mean-squared error in Sections 2.1 and 2.2.

2. For Hodges' estimator and also for the post-model-selection estimator based on a consistent model selection procedure considered in Section 2.1 the maximal (scaled) mean-squared error increases without bound as $n \to \infty$, whereas the maximal (scaled) mean-squared error of the least-squares estimator in the overall model remains bounded. Cf. Section 2.1.
3. The unboundedness of the maximal (scaled) mean-squared error is true for post-model-selection estimators based on consistent procedures more generally. Yang (2003) proves such a result in a normal linear regression framework for some sort of maximal predictive risk. A proof for the maximal [scaled] mean-squared error (in fact for the maximal [scaled] absolute bias) as considered in the present paper is given in Appendix C.[28] In contrast, the maximal (scaled) mean-squared error of a post-model-selection estimator based on a conservative (but inconsistent) procedure typically stays bounded as sample size increases (although it can substantially exceed the [scaled] mean-squared error of the least-squares estimator in the unrestricted model).[29]
4. Kempthorne (1984) has shown that in a normal linear regression model no post-model-selection estimator $\tilde{\theta}$ (including the trivial post-model-selection estimators that are based on a fixed model) dominates any other post-model-selection estimator in terms of mean-squared error of $X\tilde{\theta}$.
5. It is well known that in a normal linear regression model $Y = X\theta + \varepsilon$ with more than two regressors the least-squares estimator $(X'X)^{-1}X'Y$ is inadmissible as it is dominated by the Stein estimator (and its admissible versions). Similarly, every pretest estimator is inadmissible as shown by Sclove, Morris, and Radhakrishnan (1972). See Judge and Yancey (1986, p. 33) for more information.

Remark 4.2. That in the case of two competing models minimum AIC (and also BIC) reduces to a likelihood ratio test has been noted already by Söderström (1977) and has been rediscovered numerous times. Even in the general case there is a closer connection between model selection based on multiple testing procedures and model selection procedures based on information criteria such as AIC or BIC than is often recognized. For example, the minimum AIC or BIC method can be reexpressed as the search for that model that is not rejected in pairwise comparisons against any other competing model, where rejection occurs if the likelihood-ratio statistic (corresponding to the pairwise comparison) exceeds a critical value that is determined by the model dimensions and sample size; see Pötscher (1991, Sect. 4, Remark (ii)) for more information.

Remark 4.3. The idea that hypothesis tests give rise to consistent (model) selection procedures if the significance levels of the tests approach zero at an appropriate rate as sample size increases has already been used in Pötscher (1981,

https://doi.org/10.1017/S0266466605050036 Published online by Cambridge University Press

1983) in the context of ARMA models and in Bauer, Pötscher, and Hackl (1988) in the context of general (semi)parametric models. It has since been rediscovered numerous times, e.g., by Andrews (1986), Corradi (1999), Altissimo and Corradi (2002, 2003), and Bunea, Niu, and Wegkamp (2003), to mention a few. [The editor has informed us that in the context of a linear regression model the same idea appears also in a 1981 manuscript by Sargan, which was eventually published as Sargan, 2001.]

Remark 4.4.

1. If $\alpha_n = \alpha + \delta/\sqrt{n}$ and $\beta_n = \beta + \gamma/\sqrt{n}$ then P_{n,α_n,β_n} is contiguous w.r.t. $P_{n,\alpha,\beta}$ (and this is more generally true in any sufficiently regular parametric model). If $\breve{M}$ is an arbitrary consistent model selection procedure, i.e., satisfies $P_{n,\alpha,\beta}(\breve{M} = M_0) \to 1$ as $n \to \infty$, where $M_0 = M_0(\alpha,\beta)$ is the most parsimonious true model corresponding to (α,β), then also $P_{n,\alpha_n,\beta_n}(\breve{M} = M_0) \to 1$ as $n \to \infty$ by contiguity, and hence the post-model-selection estimator based on $\breve{M}$ coincides with the restricted estimator with P_{n,α_n,β_n} probability converging to unity if $\beta = 0$. Hence, *any* consistent model selection procedure is insensitive to deviations at least of the order $1/\sqrt{n}$. It is obvious that this argument immediately carries over to any class of sufficiently regular parametric models (except if the competing models are "well separated").
2. As a consequence of the preceding contiguity argument, in general no model selector can be uniformly consistent for the most parsimonious true model. Cf. also Corollary 2.3 in Pötscher (2002) and Corollary 3.3 in Leeb and Pötscher (2002) and observe that the estimand (i.e., the most parsimonious true model) depends discontinuously on the probability measure underlying the data generating process (except in the case where the competing models are "well separated").

Remark 4.5. Suppose that in the context of model (1) the parameter of interest is now not α but more generally a linear combination $d_1\alpha + d_2\beta$, which is estimated by $d_1\tilde{\alpha} + d_2\tilde{\beta}$, where $\tilde{\alpha}$ is the post-model-selection estimator as defined in Section 2 and the post-model-selection estimator $\tilde{\beta}$ is defined similarly, i.e., $\tilde{\beta} = \hat{\beta}(\hat{M})$. An important example is the case where the quantity of interest is a linear predictor. Then appropriate analogues to the results discussed in the present paper apply, where the rôle of ρ is now played by the correlation coefficient between $d_1\hat{\alpha}(U) + d_2\hat{\beta}(U)$ and $\hat{\beta}(U)$. See Leeb (2003a, 2003b) and Leeb and Pötscher (2003b, 2004) for a discussion in a more general framework.

Remark 4.6. We have excluded the special case $\rho_\infty = 0$ in parts of the discussion of consistent model selection procedures in Section 2.1 for the sake of simplicity. It is, however, included in the theoretical results presented in Appendix A. In the following discussion we comment on this case.

1. If $\rho = 0$ then it is easy to see that all effects from model selection disappear in the finite-sample formulas in Section 2.1. This is not surprising

https://doi.org/10.1017/S0266466605050036 Published online by Cambridge University Press

because $\rho = 0$ implies that the design matrix has orthogonal columns and hence the post-model-selection estimator $\tilde{\alpha}$ coincides with the restricted and also with the unrestricted least-squares estimator for α.

2. If only $\rho_\infty = 0$ (i.e., the columns of the design matrix are only asymptotically orthogonal), then the effects of model selection need not disappear from the asymptotic formulas; cf. Appendix A. However, inspection of the results in Appendix A shows that these effects will disappear asymptotically if ρ converges to $\rho_\infty = 0$ sufficiently fast (essentially faster than $1/c$). (In contrast, in the case of conservative model selection procedures the condition $\rho_\infty = 0$ suffices to make all effects from model selection disappear from the asymptotic formulas; cf. Section 2.2.)
3. As noted previously, in the case of an orthogonal design (i.e., $\rho = 0$) all effects from model selection on the distributional properties of $\tilde{\alpha}$ vanish. However, even for orthogonal designs, effects from model selection will nevertheless typically be present as soon as a linear combination $d_1\alpha + d_2\beta$ other than α represents the parameter of interest because then the correlation coefficient between $d_1\hat{\alpha}(U) + d_2\hat{\beta}(U)$ and $\hat{\beta}(U)$ rather than ρ governs the effects from model selection on the post-model-selection estimator; cf. Remark 4.5.

5. CONCLUSION

The distributional properties of post-model-selection estimators are quite intricate and are not properly captured by the usual pointwise large-sample analysis. The reason is lack of uniformity in the convergence of the finite-sample distributions and of associated quantities such as the bias or mean-squared error. Although it has long been known that uniformity (at least locally) w.r.t. the parameters is an important issue in asymptotic analysis, this lesson has often been forgotten in the daily practice of econometric and statistical theory where we are often content to prove pointwise asymptotic results (i.e., results that hold for each fixed true parameter value). This amnesia—and the resulting practice—fortunately has no dramatic consequences as long as only sufficiently "regular" estimators in sufficiently "regular" models are considered.[30] However, because post-model-selection estimators are quite "irregular," the uniformity issues surface here with a vengeance. Hajek's (1971, p. 153) warning,

> Especially misinformative can be those limit results that are not uniform. Then the limit may exhibit some features that are not even approximately true for any finite n . . .

thus takes on particular relevance in the context of model selection: While a pointwise asymptotic analysis paints a very misleading picture of the properties of post-model-selection estimators, an asymptotic analysis based on the fiction of a true parameter that depends on sample size provides highly accurate insights into the finite-sample properties of such estimators.

https://doi.org/10.1017/S0266466605050036 Published online by Cambridge University Press

The distinction between consistent and conservative model selection procedures is an artificial one as discussed in Section 2 and is rather a property of the embedding framework than of the model selection procedure. Viewing a model selection procedure as consistent results in a completely misleading pointwise asymptotic analysis that does not capture any of the effects of model selection that are present in finite samples. Viewing a model selection procedure as conservative (but inconsistent) results in a pointwise asymptotic analysis that captures some of the effects of model selection, although still missing others.

We would like to stress that the claim that the use of a consistent model selection procedure allows one to act as if the true model were known in advance is without any substance. In fact, any asymptotic consideration based on the so-called oracle property should not be trusted. (Somewhat ironically, consistent model selection procedures that seem not to affect the asymptotic distribution in a pointwise analysis at all exhibit stronger effects [e.g., larger maximal absolute bias or larger maximal mean-squared error] as a result of model selection in a "uniform" analysis when compared with conservative procedures.)[31] Similar warnings apply more generally to procedures that consistently choose from a finite set of alternatives (e.g., procedures that consistently decide between $I(0)$ and $I(1)$ or consistently select the number of structural breaks, etc.). Also, the claim that one can come up with a model selection procedure that can always detect the most parsimonious true model with high probability is unwarranted: However the model selection procedure is constructed, the misclassification error is always there and will be substantial for certain values of the true parameter, regardless of how large sample size is.

As shown in Section 2.3, accurate estimation of the distribution of post-model-selection estimators is intrinsically a difficult problem. In particular, it is typically impossible to estimate these distributions uniformly consistently. Similar results apply to certain shrinkage-type estimators as discussed in Section 3.

Although the discussion in this paper is set in the framework of a simple linear regression model, the issues discussed are obviously relevant much more generally. Results on post-model-selection estimators for nonlinear models and/or dependent data are given in Sen (1979), Pötscher (1991), Hjort and Claeskens (2003), and Nickl (2003).

We stress that the discussion in this paper should neither be construed as a criticism nor as an endorsement of model selection (be it consistent or conservative). In this paper we take no position on whether or not model selection is a sensible strategy. Of course, this is an important issue, but it is not the one we address here. A starting point for such a discussion could certainly be the results mentioned in Remark 4.1.

Although there is now a substantial body of literature on distributional properties of post-model-selection estimators, a proper theory of inference post model selection is only slowly emerging and is currently the subject of intensive research. We hope to be able to report on this elsewhere.

https://doi.org/10.1017/S0266466605050036 Published online by Cambridge University Press

NOTES

1. We assume throughout that at least one of the competing models is capable of correctly describing the data generating process. We do not touch upon the important question of model selection in the context of fitting only approximate models.

2. The pretest literature as summarized in Judge and Bock (1978) or Giles and Giles (1993) concentrates exclusively on second moment properties of pretest estimators and does not provide distributional results.

3. Some of the issues we raise here may not apply in the (relatively trivial) case where one selects between "well-separated" model classes, i.e., model classes that have positive minimum distance, e.g., in the Kullback–Leibler sense.

4. For example, Bunea (2004), Dufour, Pelletier, and Renault (2003, Sect. 7); Fan and Li (2001), Hall and Peixe (2003, Theorem 3), Hidalgo (2002, Theorem 3.4), and Lütkepohl (1990, p. 120) to mention a few.

5. With hindsight the second author regrets having included Lemma 1 in Pötscher (1991) at all, as this lemma seems to have contributed to popularizing the aforementioned unwarranted conclusion in the literature. Given that this lemma was included, he wishes at least that he had been more guarded in his wording in the discussion of this lemma and that he had issued a stronger warning against an uncritical use of it.

6. That is, a procedure that asymptotically selects only correct models but possibly overparameterized ones.

7. Nothing substantial changes because of this convenience assumption. The entire discussion that follows can also be given for the unknown σ^2 case. See Leeb and Pötscher (2003a) and Leeb (2003a, 2003b).

8. In fact, it would be more precise to talk about consistent (or conservative) *sequences* of model selection procedures.

9. This property of consistent model selection procedures has already been observed by Hannan and Quinn (1979, p. 191). It has since been rediscovered several times in special instances; cf. Ensor and Newton (1988, Theorem 2.1); Bunea (2004, Sect. 4).

10. Hodges' estimator (with $a = 0$ in the notation of Lehmann and Casella, 1998) is a post-model-selection estimator based on a model selection procedure that consistently chooses between an $N(0,1)$ and an $N(\theta,1)$ distribution.

11. Exceptions are Hosoya (1984), Shibata (1986), Pötscher (1991), and Kabaila (1995, 1996), who explicitly note this problem.

12. For a detailed treatment of the finite-sample properties of post-model-selection estimators in linear regression models see Leeb and Pötscher (2003a), Leeb (2003a, 2003b).

13. Slightly more general conditions under which this is true are given in Proposition A.1 in Appendix A.

14. It can be debated whether the β's giving rise to this phenomenon are justifiably viewed as "small": The phenomenon can, e.g., arise if $\beta \neq 0$ satisfies $\beta = \zeta\sigma_\beta c/\sqrt{n}$ with $|\zeta| < 1$ (cf. Proposition A.1 in Appendix A). Although such sequences of β's converge to zero by the assumption $c = o(\sqrt{n})$ maintained in Section 2.1, the "nonzeroness" of any such β can be detected with probability approaching unity by a standard test with fixed significance level or equivalently, with fixed cutoff point, and thus such β's could justifiably be classified as "far" from zero. (In more mathematical terms, $P_{n,\alpha,\beta}$ is not contiguous w.r.t. $P_{n,\alpha,0}$ for such β's.) By the way, this also nicely illustrates that the consistent model selection procedure is (not surprisingly) less powerful in detecting $\beta \neq 0$ compared with the conservative procedure with a fixed value of c, the reason being that the consistent procedure has to let the significance level of the test approach zero to asymptotically avoid choosing a model that is too large. (This loss of power is not specific to the consistent model selection procedure discussed here but is typical for consistent model selection procedures in general.)

15. In light of (2), the first term is actually the conditional density of $\sqrt{n}(\hat{\alpha}(R) - \alpha)$ given the event that the pretest does not reject multiplied by the probability of this event. Because the test

statistic is independent of $\hat{\alpha}(R)$ (Leeb and Pötscher, 2003a, Proposition 3.1), this conditional density reduces to the unconditional one. Similarly, the second term is the conditional density of $\sqrt{n}(\hat{\alpha}(U) - \alpha)$ given that the pretest rejects multiplied by the probability of this event. Because the test statistic is typically correlated with $\hat{\alpha}(U)$, the conditional density is not normal, which is reflected by the "deformation" factor.

16. A quick alternative argument showing that the convergence of the finite-sample c.d.f.s of post-model-selection estimators is typically not uniform runs as follows: Equip the space of c.d.f.s with a suitable metric (e.g., a metric that generates the topology of weak convergence). Observe that the finite-sample c.d.f.s typically depend continuously on the underlying parameters, whereas their (pointwise) limits typically are discontinuous in the underlying parameters. This shows that the convergence can not be uniform.

17. Although this fits in nicely with (5), it is not a direct consequence of (5). The crucial point here is that $P_{n,\alpha,\beta}(\hat{M} = R) = \Delta(\sqrt{n}\beta/\sigma_\beta, c)$ converges to zero exponentially fast for fixed $\beta \neq 0$; see, e.g., Lemma B.1 in Leeb and Pötscher (2003a).

18. Although this is again in line with (5) it is again not a direct consequence of (5) but follows from the exponential decay of $\Delta(\sqrt{n}\beta/\sigma_\beta, c)$ for fixed $\beta \neq 0$; cf. note 17. Furthermore, the fact that the pointwise limit of the *MSE* coincides with the asymptotic variance of the infeasible "estimator" $\hat{\alpha}(M_0)$ is not particular to the consistent model selection procedure discussed here. It is true for consistent model selection procedures in general, provided the probability of selecting an incorrect model converges to zero sufficiently fast, which is typically the case; see Nishii (1984) for some results in this direction. Of course, being only pointwise limit results, these results are subject to the criticism put forward in the present paper.

19. We could allow more generally for a sample-size-dependent c that, e.g., converges to a positive real number. See Leeb and Pötscher (2003a, Remark 6.2).

20. For a detailed treatment of the finite-sample and asymptotic properties of post-model-selection estimators based on a conservative model selection procedure see Pötscher (1991), Leeb and Pötscher (2003a), and Leeb (2003a, 2003b).

21. Similar as for consistent model selection procedures in fact all accumulation points of the model selection probabilities, the finite-sample distributions, the bias, and the mean-squared error can be characterized by a subsequence argument similar to Remark A.8; cf. also Leeb and Pötscher (2003a, Remark 4.4(i)), and Leeb (2003b, Remark 5.5).

22. Nevertheless, it is easy to see that $\tilde{\alpha}$ is consistent (cf. Pötscher, 1991, Lemma 2) and, in fact, is uniformly consistent; see Proposition B.1 in Appendix B.

23. Here the convergence of the finite-sample distribution to the asymptotic distribution is w.r.t. total variation distance.

24. Kilian (1998) claims the validity of a bootstrap procedure in the context of autoregressive models that is based on a conservative model selection procedure. Hansen (2003) makes a similar claim for a stationary bootstrap procedure in the context of a conservative model selection procedure. The preceding discussion intimates that both these claims are at least unsubstantiated.

25. Similar "impossibility" results apply to estimators of the model selection probabilities; see Leeb and Pötscher (2004) in the case of conservative procedures; for consistent procedures this argument can be easily adapted by making use of Proposition A.1.

26. The confidence interval suggested in Hjort and Claeskens (2003, p. 886) does not provide a solution to this problem. As pointed out in Remark 3.5 of Kabaila and Leeb (2004), the proposed interval (asymptotically) coincides with the classical confidence interval obtained from the overall model.

27. Although the James–Stein estimator is known to dominate the least-squares estimator in a normal linear regression model with more than two regressors, we are not aware of any similar result for the other shrinkage-type estimators mentioned earlier. (In fact, for some it is known that they do not dominate the least-squares estimator.)

28. This proof seems to be somewhat simpler than Yang's proof and has the advantage of also covering nonnormally distributed errors. It should easily extend to Yang's framework, but we do not pursue this here.

29. The fact that the maximal (scaled) mean-squared error remains bounded for conservative procedures is sometimes billed as "minimax rate optimality" of the procedure (see, e.g., Yang, 2003, and the references given there). Given that this "optimality" property is typically shared by any post-model-selection estimator based on a conservative procedure (including the procedure that always selects the overall model), this property does not seem to carry much weight here.

30. The reason is that the asymptotic properties of such estimators typically are then in fact "automatically" uniform, at least locally.

31. This is not surprising. For the particular model selection procedure considered here it is obvious that a larger value of the cutoff point c gives more "weight" to the restricted model, which results in a larger maximal absolute bias.

REFERENCES

Ahmed, S.E. & A.K. Basu (2000) Least squares, preliminary test and Stein-type estimation in general vector AR(p) models. *Statistica Neerlandica* 54, 47–66.

Altissimo, F. & V. Corradi (2002) Bounds for inference with nuisance parameters present only under the alternative. *Econometrics Journal* 5, 494–519.

Altissimo, F. & V. Corradi (2003) Strong rules for detecting the numbers of breaks in a time series. *Journal of Econometrics* 117, 207–244.

Andrews, D.W.K. (1986) Complete consistency: A testing analogue of estimator consistency. *Review of Economic Studies* 53, 263–269.

Bauer, P., B.M. Pötscher, & P. Hackl (1988) Model selection by multiple test procedures. *Statistics* 19, 39–44.

Bunea, F. (2004) Consistent covariate selection and post model selection inference in semiparametric regression. *Annals of Statistics* 32, 898–927.

Bunea, F., X. Niu, & M.H. Wegkamp (2003) The Consistency of the FDR Estimator. Working paper, Department of Statistics, Florida State University at Tallahassee.

Chen, S.S., D.L. Donoho, & M.A. Saunders (1998) Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing* 20, 33–61.

Corradi, V. (1999) Deciding between $I(0)$ and $I(1)$ via FLIL-based bounds. *Econometric Theory* 15, 643–663.

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

Dijkstra, T.K. & J.H. Veldkamp (1988) Data-driven selection of regressors and the bootstrap. *Lecture Notes in Economics and Mathematical Systems* 307, 17–38.

Dufour, J.M., D. Pelletier, & E. Renault (2003) Short run and long run causality in time series: Inference. *Journal of Econometrics* (forthcoming).

Dukić, V.M. & E.A Peña (2002) Estimation after Model Selection in a Gaussian Model. Manuscript, Department of Statistics, University of Chicago.

Ensor, K.B. & H.J. Newton (1988) The effect of order estimation on estimating the peak frequency of an autoregressive spectral density. *Biometrika* 75, 587–589.

Fan, J. & R. Li (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96, 1348–1360.

Frank, I.E. & J.H. Friedman (1993) A statistical view of some chemometrics regression tools (with discussion). *Technometrics* 35, 109–148.

Giles, J.A. & D.E.A. Giles (1993) Pre-test estimation and testing in econometrics: Recent developments. *Journal of Economic Surveys* 7, 145–197.

Hajek, J. (1971) Limiting properties of likelihoods and inference. In V.P. Godambe & D.A. Sprott (eds.), *Foundations of Statistical Inference: Proceedings of the Symposium on the Foundations of Statistical Inference, University of Waterloo, Ontario, March 31–April 9, 1970*, pp. 142–159. Holt, Rinehart and Winston.

Hajek, J. & Z. Sidak (1967) *Theory of Rank Tests*. Academic Press.

Hall, A.R. & F.P.M. Peixe (2003) A consistent method for the selection of relevant instruments. *Econometric Reviews* 22, 269–287.

Hannan, E.J. & B.G. Quinn (1979) The determination of the order of an autoregression. *Journal of the Royal Statistical Society, Series B* 41, 190–195.

Hansen, P.R. (2003) Regression Analysis with Many Specifications: A Bootstrap Method for Robust Inference. Working paper, Department of Economics, Brown University.

Hidalgo, J. (2002) Consistent order selection with strongly dependent data and its application to efficient estimation. *Journal of Econometrics* 110, 213–239.

Hjort, N.L. & G. Claeskens (2003) Frequentist model average estimators. *Journal of the American Statistical Association* 98, 879–899.

Hosoya, Y. (1984) Information criteria and tests for time series models. In O.D. Anderson (ed.), *Time Series Analysis: Theory and Practice*, vol. 5, pp. 39–52. North-Holland.

Judge, G.G. & M.E. Bock (1978) *The Statistical Implications of Pre-test and Stein-Rule Estimators in Econometrics*. North-Holland.

Judge, G.G. & T.A. Yancey (1986) *Improved Methods of Inference in Econometrics*. North-Holland.

Kabaila, P. (1995) The effect of model selection on confidence regions and prediction regions. *Econometric Theory* 11, 537–549.

Kabaila, P. (1996) The evaluation of model selection criteria: Pointwise limits in the parameter space. In D.L. Dowe, K.B. Korb, & J.J. Oliver (eds.), *Information, Statistics and Induction in Science*, pp. 114–118. World Scientific.

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

Kabaila, P. & H. Leeb (2004) On the Large-Sample Minimal Coverage Probability of Confidence Intervals after Model Selection. Working paper, Department of Statistics, Yale University.

Kapetanios, G. (2001) Incorporating lag order selection uncertainty in parameter inference for AR models. *Economics Letters* 72, 137–144.

Kempthorne, P.J. (1984) Admissible variable-selection procedures when fitting regression models by least squares for prediction. *Biometrika* 71, 593–597.

Kilian, L. (1998) Accounting for lag order uncertainty in autoregressions: The endogenous lag order bootstrap algorithm. *Journal of Time Series Analysis* 19, 531–548.

Knight, K. (1999) Epi-convergence in Distribution and Stochastic Equi-semicontinuity. Working paper, Department of Statistics, University of Toronto.

Knight, K. & W. Fu (2000) Asymptotics of lasso-type estimators. *Annals of Statistics* 28, 1356–1378.

Koul, H.L. & W. Wang (1984) Local asymptotic normality of randomly censored linear regression model. *Statistics & Decisions*, supplement 1, 17–30.

Kulperger, R.J. & S.E. Ahmed (1992) A bootstrap theorem for a preliminary test estimator. *Communications in Statistics: Theory and Methods* 21, 2071–2082.

Leeb, H. (2003a) The distribution of a linear predictor after model selection: Conditional finite-sample distributions and asymptotic approximations. *Journal of Statistical Planning and Inference* (forthcoming).

Leeb, H. (2003b) The Distribution of a Linear Predictor after Model Selection: Unconditional Finite-Sample Distributions and Asymptotic Approximations. Working paper, Department of Statistics, University of Vienna.

Leeb, H. & B.M. Pötscher (2002) Performance Limits for Estimators of the Risk or Distribution of Shrinkage-Type Estimators, and Some General Lower Risk-Bound Results. Working paper, Department of Statistics, University of Vienna.

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

Leeb, H. & B.M. Pötscher (2003b) Can One Estimate the Conditional Distribution of Post-Model-Selection Estimators? Working paper, Department of Statistics, University of Vienna. (Also available as Cowles Foundation Discussion paper 1444.)

Leeb, H. & B.M. Pötscher (2004) Can One Estimate the Unconditional Distribution of Post-Model-Selection Estimators? Manuscript, Department of Statistics, Yale University.

Lehmann, E.L. & G. Casella (1998) *Theory of Point Estimation*. Springer Texts in Statistics. Springer-Verlag.

Lütkepohl, H. (1990) Asymptotic distributions of impulse response functions and forecast error variance decompositions of vector autoregressive models. *Review of Economics and Statistics* 72, 116–125.

Magnus, J.R. (1999) The traditional pretest estimator. *Teoriya Veroyatnost. i Primenen.* 44, 401–418; translation in *Theory of Probability and Its Applications* 44 (2000), 293–308.

Nickl, R. (2003) *Asymptotic Distribution Theory of Post-Model-Selection Maximum Likelihood Estimators*. Master's thesis, Department of Statistics, University of Vienna.

Nishii, R. (1984) Asymptotic properties of criteria for selection of variables in multiple regression. *Annals of Statistics* 12, 758–765.

Phillips, P.C.B. (2005) Automated discovery in econometrics. *Econometric Theory* (this issue).

Pötscher, B.M. (1981) Order Estimation in ARMA-Models by Lagrangian Multiplier Tests. Research report 5, Department of Econometrics and Operations Research, University of Technology, Vienna.

Pötscher, B.M. (1983) Order estimation in ARMA-models by Lagrangian multiplier tests. *Annals of Statistics* 11, 872–885.

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

Pötscher, B.M. (1995) Comment on "The effect of model selection on confidence regions and prediction regions." *Econometric Theory* 11, 550–559.

Pötscher, B.M. (2002) Lower risk bounds and properties of confidence sets for ill-posed estimation problems with applications to spectral density and persistence estimation, unit roots, and estimation of long memory parameters. *Econometrica* 70, 1035–1065.

Pötscher, B.M. & A.J. Novak (1998) The distribution of estimators after model selection: Large and small sample results. *Journal of Statistical Computation and Simulation* 60, 19–56.

Rao, C.R. & Y. Wu (2001) On model selection. *IMS Lecture Notes Monograph Series* 38, 1–57.

Sargan, D.J. (2001) The choice between sets of regressors. *Econometric Reviews* 20, 171–186.

Sclove, S.L., C. Morris, & R. Radhakrishnan (1972) Non-optimality of preliminary-test estimators for the mean of a multivariate normal distribution. *Annals of Mathematical Statistics* 43, 1481–1490.

Sen, P.K (1979) Asymptotic properties of maximum likelihood estimators based on conditional specification. *Annals of Statistics* 7, 1019–1033.

Sen, P.K & A.K.M.E. Saleh (1987) On preliminary test and shrinkage *M*-estimation in linear models. *Annals of Statistics* 15, 1580–1592.

Shibata, R. (1986) Consistency of model selection and parameter estimation. *Journal of Applied Probability*, special volume 23A, 127–141.

Söderström, T. (1977) On model structure testing in system identification. *International Journal of Control* 26, 1–18.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* 58, 267–288.

Yang, Y. (2003) Can the Strengths of AIC and BIC Be Shared? Working paper, Department of Statistics, Iowa State University.

APPENDIX A: ASYMPTOTIC RESULTS FOR CONSISTENT MODEL SELECTION PROCEDURES

In this Appendix we provide propositions that together with Remark A.8, which follows, characterize all possible limits (more precisely, all accumulation points) of the model selection probabilities, the finite-sample distribution, the (scaled) bias, and the (scaled) mean-squared error of the post-model-selection estimator based on a consistent model selection procedure under arbitrary sequences of parameters (α_n, β_n). Recall that these quantities do not depend on α and hence the behavior of α will not enter

https://doi.org/10.1017/S0266466605050036 Published online by Cambridge University Press

the results in the sequel. In the following discussion we consider the linear regression model (1) under the assumptions of Section 2. Furthermore, we assume as in Section 2.1 that $c \to \infty$ and $c/\sqrt{n} \to 0$ as $n \to \infty$.

PROPOSITION A.1. *Let (α_n, β_n) be an arbitrary sequence of values for the regression parameters in (1).*

1. *Suppose $\sqrt{n}\beta_n/(\sigma_\beta c) \to \zeta$, $|\zeta| < 1$, as $n \to \infty$. Then $\lim_{n\to\infty} P_{n,\alpha_n,\beta_n}(\hat{M} = R) = 1$.*
2. *Suppose $\sqrt{n}\beta_n/(\sigma_\beta c) \to \zeta$, $1 < |\zeta| \leq \infty$, as $n \to \infty$. Then $\lim_{n\to\infty} P_{n,\alpha_n,\beta_n}(\hat{M} = R) = 0$.*
3. *Suppose $\sqrt{n}\beta_n/(\sigma_\beta c) \to 1$ and $c - \sqrt{n}\beta_n/\sigma_\beta \to r$ for some $r \in \mathbb{R} \cup \{-\infty, \infty\}$ as $n \to \infty$. Then $\lim_{n\to\infty} P_{n,\alpha_n,\beta_n}(\hat{M} = R) = \Phi(r)$.*
4. *Suppose $\sqrt{n}\beta_n/(\sigma_\beta c) \to -1$ and $c + \sqrt{n}\beta_n/\sigma_\beta \to s$ for some $s \in \mathbb{R} \cup \{-\infty, \infty\}$ as $n \to \infty$. Then $\lim_{n\to\infty} P_{n,\alpha_n,\beta_n}(\hat{M} = R) = \Phi(s)$.*

Proof. From (3) we have

$$P_{n,\alpha_n,\beta_n}(\hat{M} = R) = \Phi(c - \sqrt{n}\beta_n/\sigma_\beta) - \Phi(-c - \sqrt{n}\beta_n/\sigma_\beta).$$

Observe that $\Phi(c - \sqrt{n}\beta_n/\sigma_\beta) = \Phi(c(1 - \sqrt{n}\beta_n/(\sigma_\beta c)))$ and $\Phi(-c - \sqrt{n}\beta_n/\sigma_\beta) = \Phi(c(-1 - \sqrt{n}\beta_n/(\sigma_\beta c)))$. The first two claims then follow immediately. The third claim follows because then $\Phi(c - \sqrt{n}\beta_n/\sigma_\beta)$ trivially converges to $\Phi(r)$, whereas $\Phi(-c - \sqrt{n}\beta_n/\sigma_\beta) = \Phi(c(-1 - \sqrt{n}\beta_n/(\sigma_\beta c)))$ converges to zero. The fourth claim is proved analogously. ■

The next proposition describes the possible limiting behavior of the finite-sample distribution of the post-model-selection estimator, which is somewhat complex. It turns out that the limit can, e.g., be point-mass at (plus or minus) infinity, or a convex combination of such a point-mass with a "deformed" normal distribution, or a convex combination of a normal distribution with a "deformed" normal. Let $G_{n,\alpha,\beta}(t)$ denote the cumulative distribution function corresponding to the density $g_{n,\alpha,\beta}(u)$ of $\sqrt{n}(\tilde{\alpha} - \alpha)$. Also recall that convergence in total variation of a sequence of absolutely continuous c.d.f.s on the real line is equivalent to convergence of the densities in the L^1-sense.

PROPOSITION A.2. *Let (α_n, β_n) be an arbitrary sequence of values for the regression parameters in (1).*

1. *Suppose that (i) $\sqrt{n}\beta_n/(\sigma_\beta c) \to \zeta$, $|\zeta| < 1$, or that (ii) $\sqrt{n}\beta_n/(\sigma_\beta c) \to 1$, $c - \sqrt{n}\beta_n/\sigma_\beta \to \infty$, or that (iii) $\sqrt{n}\beta_n/(\sigma_\beta c) \to -1$, $c + \sqrt{n}\beta_n/\sigma_\beta \to \infty$ as $n \to \infty$. Assume furthermore that $\sqrt{n}\beta_n(\rho/\sigma_\beta) \to \chi$ for some $\chi \in \mathbb{R} \cup \{-\infty, \infty\}$ as $n \to \infty$. If $\chi = -\infty$, then $G_{n,\alpha_n,\beta_n}(t)$ converges to 0 for every $t \in \mathbb{R}$; i.e., $\sqrt{n}(\tilde{\alpha} - \alpha_n)$ converges to ∞ in P_{n,α_n,β_n} probability. If $\chi = \infty$, then $G_{n,\alpha_n,\beta_n}(t)$ converges to 1 for every $t \in \mathbb{R}$; i.e., $\sqrt{n}(\tilde{\alpha} - \alpha_n)$ converges to $-\infty$ in P_{n,α_n,β_n} probability. If $|\chi| < \infty$, then $G_{n,\alpha_n,\beta_n}(t)$ converges to $\Phi((1 - \rho_\infty^2)^{-1/2} \times (t/\sigma_{\alpha,\infty} + \chi))$ in total variation distance; in fact, $g_{n,\alpha_n,\beta_n}(u)$ converges to $\sigma_{\alpha,\infty}^{-1}(1 - \rho_\infty^2)^{-1/2}\phi((1 - \rho_\infty^2)^{-1/2}(u/\sigma_{\alpha,\infty} + \chi))$ pointwise and hence in the L^1 sense.*
2. *Suppose that (i) $\sqrt{n}\beta_n/(\sigma_\beta c) \to \zeta$, $1 < |\zeta| \leq \infty$, or that (ii) $\sqrt{n}\beta_n/(\sigma_\beta c) \to 1$, $c - \sqrt{n}\beta_n/\sigma_\beta \to -\infty$, or that (iii) $\sqrt{n}\beta_n/(\sigma_\beta c) \to -1$, $c + \sqrt{n}\beta_n/\sigma_\beta \to -\infty$ as $n \to \infty$. Then $G_{n,\alpha_n,\beta_n}(t)$ converges to $\Phi(t/\sigma_{\alpha,\infty})$ in the total variation distance;*

https://doi.org/10.1017/S0266466605050036 Published online by Cambridge University Press

in fact, $g_{n,\alpha_n,\beta_n}(u)$ converges to $\sigma_{\alpha,\infty}^{-1}\phi(u/\sigma_{\alpha,\infty})$ pointwise and hence in the L^1 sense.

3. *Suppose that $\sqrt{n}\beta_n/(\sigma_\beta c) \to 1$, $c - \sqrt{n}\beta_n/\sigma_\beta \to r$ for some $r \in \mathbb{R}$ and $\sqrt{n}\beta_n(\rho/\sigma_\beta) \to \chi$ for some $\chi \in \mathbb{R} \cup \{-\infty,\infty\}$ as $n \to \infty$. If $|\chi| = \infty$, then $G_{n,\alpha_n,\beta_n}(t)$ converges to*

$$\Phi(r)\mathbf{1}(\chi > 0) + \int_{-\infty}^{t} \sigma_{\alpha,\infty}^{-1}\phi(u/\sigma_{\alpha,\infty})\Phi((1-\rho_\infty^2)^{-1/2}(-r + \rho_\infty\sigma_{\alpha,\infty}^{-1}u))\,du \tag{A.1}$$

for every $t \in \mathbb{R}$. The limit is a convex combination of pointmass at $\operatorname{sign}(-\chi)\infty$ and a c.d.f. with density given by $1/(1-\Phi(r))$ times the integrand in the preceding display, the weights in the convex combination given by $\Phi(r)$ and $1-\Phi(r)$, respectively. If $|\chi| < \infty$, then $G_{n,\alpha_n,\beta_n}(t)$ converges to

$$\Phi(r)\int_{-\infty}^{t} \sigma_{\alpha,\infty}^{-1}(1-\rho_\infty^2)^{-1/2}\phi((1-\rho_\infty^2)^{-1/2}(u/\sigma_{\alpha,\infty} + \chi))\,du$$

$$+ \int_{-\infty}^{t} \sigma_{\alpha,\infty}^{-1}\phi(u/\sigma_{\alpha,\infty})\Phi((1-\rho_\infty^2)^{-1/2}(-r + \rho_\infty\sigma_{\alpha,\infty}^{-1}u))\,du$$

for every $t \in \mathbb{R}$.

4. *Suppose $\sqrt{n}\beta_n/(\sigma_\beta c) \to -1$ and $c + \sqrt{n}\beta_n/\sigma_\beta \to s$ for some $s \in \mathbb{R}$, and $\sqrt{n}\beta_n(\rho/\sigma_\beta) \to \chi$ for some $\chi \in \mathbb{R} \cup \{-\infty,\infty\}$ as $n \to \infty$. If $|\chi| = \infty$, then $G_{n,\alpha_n,\beta_n}(t)$ converges to*

$$\Phi(s)\mathbf{1}(\chi > 0) + \int_{-\infty}^{t} \sigma_{\alpha,\infty}^{-1}\phi(u/\sigma_{\alpha,\infty})(1-\Phi((1-\rho_\infty^2)^{-1/2}(s + \rho_\infty\sigma_{\alpha,\infty}^{-1}u)))\,du \tag{A.2}$$

for every $t \in \mathbb{R}$. The limit is a convex combination of pointmass at $\operatorname{sign}(-\chi)\infty$ and a c.d.f. with density given by $1/(1-\Phi(s))$ times the integrand in the preceding display, the weights in the convex combination given by $\Phi(s)$ and $1-\Phi(s)$, respectively. If $|\chi| < \infty$, then $G_{n,\alpha_n,\beta_n}(t)$ converges to

$$\Phi(s)\int_{-\infty}^{t} \sigma_{\alpha,\infty}^{-1}(1-\rho_\infty^2)^{-1/2}\phi((1-\rho_\infty^2)^{-1/2}(u/\sigma_{\alpha,\infty} + \chi))\,du$$

$$+ \int_{-\infty}^{t} \sigma_{\alpha,\infty}^{-1}\phi(u/\sigma_{\alpha,\infty})(1-\Phi((1-\rho_\infty^2)^{-1/2}(s + \rho_\infty\sigma_{\alpha,\infty}^{-1}u)))\,du$$

for every $t \in \mathbb{R}$.

Proof. In view of (2) we can write the density $g_{n,\alpha,\beta}$ as

$$\begin{aligned} g_{n,\alpha,\beta}(u) &= g_{n,\alpha,\beta}(u|R)P_{n,\alpha,\beta}(\hat{M} = R) + g_{n,\alpha,\beta}(u|U)P_{n,\alpha,\beta}(\hat{M} = U) \\ &= g_{n,\alpha,\beta}(u|R)\Delta(\sqrt{n}\beta/\sigma_\beta, c) + g_{n,\alpha,\beta}(u|U)(1-\Delta(\sqrt{n}\beta/\sigma_\beta, c)), \end{aligned} \tag{A.3}$$

https://doi.org/10.1017/S0266466605050036 Published online by Cambridge University Press

where $g_{n,\alpha,\beta}(u|R)$ is the conditional density of $\sqrt{n}(\tilde{\alpha} - \alpha)$ given that $\hat{M} = R$ and $g_{n,\alpha,\beta}(u|U)$ is defined analogously. As mentioned in note 15,

$$g_{n,\alpha,\beta}(u|R) = \sigma_\alpha^{-1}(1-\rho^2)^{-1/2}\phi(u(1-\rho^2)^{-1/2}/\sigma_\alpha + \rho(1-\rho^2)^{-1/2}\sqrt{n}\beta/\sigma_\beta), \quad \textbf{(A.4)}$$

$$g_{n,\alpha,\beta}(u|U) = \sigma_\alpha^{-1}\left[\left(1 - \Delta\left(\frac{\sqrt{n}\beta/\sigma_\beta + \rho u/\sigma_\alpha}{\sqrt{1-\rho^2}}, \frac{c}{\sqrt{1-\rho^2}}\right)\right)\Big/ (1 - \Delta(\sqrt{n}\beta/\sigma_\beta, c))\right]\phi(u/\sigma_\alpha). \quad \textbf{(A.5)}$$

To prove part 1 replace (α, β) by (α_n, β_n) in the preceding formulas and observe that under the assumptions of this part of the proposition the probability $P_{n,\alpha_n,\beta_n}(\hat{M} = R)$ converges to unity (Proposition A.1) and hence the contribution to the total probability mass by the second term on the far r.h.s. of (A.3) vanishes asymptotically. It hence suffices to consider the first term only. Now $\sqrt{n}\beta_n(\rho/\sigma_\beta) \to \chi$ by assumption. Furthermore, $\rho \to \rho_\infty \neq \pm 1$ (because Q was assumed to be positive definite), and $\sigma_\alpha \to \sigma_{\alpha,\infty} > 0$. If $\chi = \pm\infty$, inspection of (A.4) immediately shows that the total probability mass of $\sqrt{n}(\tilde{\alpha} - \alpha_n)$ escapes to $\mp\infty$. If χ is finite, inspection of (A.4) reveals that the conditional density $g_{n,\alpha_n,\beta_n}(u|R)$ converges to $\sigma_{\alpha,\infty}^{-1}(1-\rho_\infty^2)^{-1/2}\phi((1-\rho_\infty^2)^{-1/2} \times (u/\sigma_{\alpha,\infty} + \chi))$ for every $u \in \mathbb{R}$. Because the limit function is a density again, convergence takes place in the L^1 sense in view of Scheffé's theorem. This establishes convergence of the corresponding c.d.f. in the total variation distance.

To prove part 2 again replace (α, β) by (α_n, β_n) in the preceding formulas and observe that under the assumptions of this part of the proposition the probability $P_{n,\alpha_n,\beta_n}(\hat{M} = R)$ converges to zero (Proposition A.1) and hence the contribution to the total probability mass by the first term on the far r.h.s. of (A.3) vanishes asymptotically. It hence suffices to consider the second term only. Now, $\rho \to \rho_\infty \neq \pm 1$, and $\sigma_\alpha \to \sigma_{\alpha,\infty} > 0$. Inspection of (A.5) then immediately shows that $g_{n,\alpha_n,\beta_n}(u|U)$ converges to $\sigma_{\alpha,\infty}^{-1}\phi(u/\sigma_{\alpha,\infty})$ for every $u \in \mathbb{R}$.

To prove part 3 observe that under the assumptions of this part of the proposition $P_{n,\alpha_n,\beta_n}(\hat{M} = R) \to \Phi(r) > 0$ and $P_{n,\alpha_n,\beta_n}(\hat{M} = U) \to 1 - \Phi(r) > 0$ hold. The proof that the total probability mass of $g_{n,\alpha_n,\beta_n}(u|R)$ escapes to $\mp\infty$ if $\chi = \pm\infty$ is exactly the same as in the proof of part 1. In the case that χ is finite, the same argument as in the proof of part 1 shows that $g_{n,\alpha_n,\beta_n}(u|R)$ converges to $\sigma_{\alpha,\infty}^{-1}(1-\rho_\infty^2)^{-1/2} \times \phi((1-\rho_\infty^2)^{-1/2}(u/\sigma_{\alpha,\infty} + \chi))$ for every $u \in \mathbb{R}$ and in L^1. Now regarding $g_{n,\alpha_n,\beta_n}(u|U)$ inspection of (A.5) shows that this density converges to $\sigma_{\alpha,\infty}^{-1}\phi(u/\sigma_{\alpha,\infty})\Phi((1-\rho_\infty^2)^{-1/2}(-r + \rho_\infty\sigma_{\alpha,\infty}^{-1}u))/(1-\Phi(r))$ for every $u \in \mathbb{R}$. Because this limit is a probability density as is readily seen, the convergence is also in L^1 by an application of Scheffé's theorem.

The proof of part 4 is completely analogous to the proof of part 3. ■

Remark A.3. In the important case where $\rho_\infty \neq 0$ the preceding results simplify somewhat: If $\rho_\infty \neq 0$ and $\zeta = \lim_{n\to\infty}\sqrt{n}\beta_n/(\sigma_\beta c) \neq 0$ in part 1 of the proposition, then necessarily $\chi = \text{sign}(\rho_\infty\zeta)\infty$; i.e., $\sqrt{n}(\tilde{\alpha} - \alpha_n)$ always converges to $\pm\infty$ in probability. If $\rho_\infty \neq 0$ in part 3 of the proposition, then necessarily $\chi = \text{sign}(\rho_\infty)\infty$; i.e., only the distribution (A.1) can arise. If $\rho_\infty \neq 0$ in part 4 of the proposition, then necessarily $\chi = \text{sign}(-\rho_\infty)\infty$; i.e., only the distribution (A.2) can arise.

https://doi.org/10.1017/S0266466605050036 Published online by Cambridge University Press

PROPOSITION A.4. *Let (α_n, β_n) be an arbitrary sequence of values for the regression parameters in (1).*

1. *Suppose that $\sqrt{n}\beta_n/(\sigma_\beta c) \to \zeta$, $|\zeta| < 1$, and that $\sqrt{n}\beta_n(\rho/\sigma_\beta) \to \chi$ for some $\chi \in \mathbb{R} \cup \{-\infty, \infty\}$ as $n \to \infty$. Then Bias $\to -\sigma_{\alpha,\infty}\chi$.*
2. *Suppose that $\sqrt{n}\beta_n/(\sigma_\beta c) \to \zeta$, $1 < |\zeta| \leq \infty$ as $n \to \infty$. Then Bias $\to 0$.*
3. *Suppose that $\sqrt{n}\beta_n/(\sigma_\beta c) \to 1$, $c - \sqrt{n}\beta_n/\sigma_\beta \to r$ for some $r \in \mathbb{R} \cup \{-\infty, \infty\}$, and $\sqrt{n}\beta_n(\rho/\sigma_\beta) \to \chi$ for some $\chi \in \mathbb{R} \cup \{-\infty, \infty\}$ as $n \to \infty$. If $r > -\infty$, or if $r = -\infty$ but χ is finite, then Bias $\to -\sigma_{\alpha,\infty}\chi\Phi(r) + \sigma_{\alpha,\infty}\rho_\infty\phi(r)$. If $r = -\infty$ and $|\chi| = \infty$, then Bias $\to -\sigma_{\alpha,\infty} \lim_{n\to\infty} \rho\sqrt{n}\beta_n(\sqrt{n}\beta_n - c\sigma_\beta)^{-1}\phi((\sqrt{n}\beta_n - c\sigma_\beta)/\sigma_\beta)$ provided this limit exists.*
4. *Suppose that $\sqrt{n}\beta_n/(\sigma_\beta c) \to -1$, $c + \sqrt{n}\beta_n/\sigma_\beta \to s$ for some $s \in \mathbb{R} \cup \{-\infty, \infty\}$, and $\sqrt{n}\beta_n(\rho/\sigma_\beta) \to \chi$ for some $\chi \in \mathbb{R} \cup \{-\infty, \infty\}$ as $n \to \infty$. If $s > -\infty$, or if $s = -\infty$ but χ is finite, then Bias $\to -\sigma_{\alpha,\infty}\chi\Phi(s) - \sigma_{\alpha,\infty}\rho_\infty\phi(s)$. If $s = -\infty$ and $|\chi| = \infty$, then Bias $\to \sigma_{\alpha,\infty} \lim_{n\to\infty} \rho\sqrt{n}\beta_n(\sqrt{n}\beta_n + c\sigma_\beta)^{-1} \times \phi((\sqrt{n}\beta_n + c\sigma_\beta)/\sigma_\beta)$ provided this limit exists.*

Proof. Under the assumptions of part 1 of the proposition $P_{n,\alpha_n,\beta_n}(\hat{M} = R) = \Delta(\sqrt{n}\beta_n/\sigma_\beta, c)$ converges to unity by Proposition A.1. Hence the first term in (11) converges to $-\sigma_{\alpha,\infty}\chi$. Because $\rho \to \rho_\infty$, $\sigma_\alpha \to \sigma_{\alpha,\infty}$, and because $\phi(\sqrt{n}\beta_n/\sigma_\beta - c)$ and also $\phi(\sqrt{n}\beta_n/\sigma_\beta + c)$ converge to zero, the second and third term in (11) go to zero, completing the proof of part 1.

To prove part 2 observe that the second and third term in (11) again converge to zero. Now, $\Delta(\sqrt{n}\beta_n/\sigma_\beta, c)$ converges to zero by Proposition A.1, whereas $\sqrt{n}\beta_n/\sigma_\beta$ diverges to $\pm\infty$. Because $\Delta(\cdot,\cdot)$ is symmetric in its first argument, we may assume that ζ is positive. Applying Lemma B.1 in Leeb and Pötscher (2003a), the limit of the first term in (11) is then readily seen to be zero.

We next prove part 3. From Proposition A.1 we see that $\Delta(\sqrt{n}\beta_n/\sigma_\beta, c)$ converges to $\Phi(r)$. Furthermore, $-\rho\sigma_\alpha\sqrt{n}\beta_n/\sigma_\beta$ converges to $-\sigma_{\alpha,\infty}\chi$ (which may be infinite). This shows that the first term in (11) converges to $-\sigma_{\alpha,\infty}\chi\Phi(r)$ provided χ is finite or $\Phi(r)$ is positive. The second term obviously converges to $\rho_\infty\sigma_{\alpha,\infty}\phi(-r) = \rho_\infty\sigma_{\alpha,\infty}\phi(r)$ (which is zero in case $r = -\infty$), whereas the third term goes to zero. If χ is infinite and $\Phi(r)$ is zero (i.e., if $r = -\infty$), Lemma B.1 in Leeb and Pötscher (2003a) shows that the first term in (11) converges to the claimed limit.

Part 4 is proved analogously to part 3. ■

Remark A.5. In the important case where $\rho_\infty \neq 0$ the following simplifications arise: If $\rho_\infty \neq 0$ and $\zeta \neq 0$ in part 1 of the proposition, then necessarily $\chi = \text{sign}(\rho_\infty\zeta)\infty$. If $\rho_\infty \neq 0$ in part 3 of the proposition, then necessarily $\chi = \text{sign}(\rho_\infty)\infty$. If $\rho_\infty \neq 0$ in part 4 of the proposition, then necessarily $\chi = \text{sign}(-\rho_\infty)\infty$.

PROPOSITION A.6. *Let (α_n, β_n) be an arbitrary sequence of values for the regression parameters in (1).*

1. *Suppose that $\sqrt{n}\beta_n/(\sigma_\beta c) \to \zeta$, $|\zeta| < 1$, and that $\sqrt{n}\beta_n(\rho/\sigma_\beta) \to \chi$ for some $\chi \in \mathbb{R} \cup \{-\infty, \infty\}$ as $n \to \infty$. Then MSE $\to \sigma^2_{\alpha,\infty}(1 - \rho^2_\infty + \chi^2)$, which is infinite if $|\chi| = \infty$.*
2. *Suppose that $\sqrt{n}\beta_n/(\sigma_\beta c) \to \zeta$, $1 < |\zeta| \leq \infty$ as $n \to \infty$. Then MSE $\to \sigma^2_{\alpha,\infty}$.*

https://doi.org/10.1017/S0266466605050036 Published online by Cambridge University Press

3. *Suppose that* $\sqrt{n}\beta_n/(\sigma_\beta c) \to 1$, $c - \sqrt{n}\beta_n/\sigma_\beta \to r$ *for some* $r \in \mathbb{R} \cup \{-\infty,\infty\}$, *and* $\sqrt{n}\beta_n(\rho/\sigma_\beta) \to \chi$ *for some* $\chi \in \mathbb{R} \cup \{-\infty,\infty\}$ *as* $n \to \infty$. *Then* $MSE \to \sigma^2_{\alpha,\infty}(1 + \rho^2_\infty r\phi(r) - \rho^2_\infty \Phi(r) + \chi^2\Phi(r))$ *if* $r > -\infty$, *or if* $r = -\infty$ *but* χ *is finite (with the convention that* $r\phi(r) = 0$ *if* $r = \pm\infty$*). If* $r = -\infty$ *and* $|\chi| = \infty$, *then* $MSE \to \sigma^2_{\alpha,\infty}[1 + \lim_{n\to\infty} \rho^2\sigma_\beta^{-1} n\beta_n^2(\sqrt{n}\beta_n - c\sigma_\beta)^{-1}\phi((\sqrt{n}\beta_n - c\sigma_\beta)/\sigma_\beta)]$ *provided this limit exists.*
4. *Suppose that* $\sqrt{n}\beta_n/(\sigma_\beta c) \to -1$, $c + \sqrt{n}\beta_n/\sigma_\beta \to s$ *for some* $s \in \mathbb{R} \cup \{-\infty,\infty\}$, *and* $\sqrt{n}\beta_n(\rho/\sigma_\beta) \to \chi$ *for some* $\chi \in \mathbb{R} \cup \{-\infty,\infty\}$ *as* $n \to \infty$. *Then* $MSE \to \sigma^2_{\alpha,\infty}(1 + \rho^2_\infty s\phi(s) - \rho^2_\infty \Phi(s) + \chi^2\Phi(s))$ *if* $s > -\infty$, *or if* $s = -\infty$ *but* χ *is finite (with the convention that* $s\phi(s) = 0$ *if* $s = \pm\infty$*). If* $s = -\infty$ *and* $|\chi| = \infty$, *then* $MSE \to \sigma^2_{\alpha,\infty}[1 - \lim_{n\to\infty} \rho^2\sigma_\beta^{-1} n\beta_n^2(\sqrt{n}\beta_n + c\sigma_\beta)^{-1}\phi((\sqrt{n}\beta_n + c\sigma_\beta)/\sigma_\beta)]$ *provided this limit exists.*

Proof. Under the assumptions of part 1 of the proposition the terms in (12) involving the standard normal density ϕ are readily seen to converge to zero. By Proposition A.1, $\Phi(c - \sqrt{n}\beta_n/\sigma_\beta) - \Phi(-c - \sqrt{n}\beta_n/\sigma_\beta)$ converges to unity. Consequently, $MSE \to \sigma^2_{\alpha,\infty}(1 - \rho^2_\infty + \chi^2)$.

To prove part 2, observe that the terms in (12) involving the standard normal density ϕ again converge to zero and that $\Phi(c - \sqrt{n}\beta_n/\sigma_\beta) - \Phi(-c - \sqrt{n}\beta_n/\sigma_\beta)$ converges to zero by Proposition A.1. Hence we only need to show that $n(\beta_n/\sigma_\beta)^2[\Phi(c - \sqrt{n}\beta_n/\sigma_\beta) - \Phi(-c - \sqrt{n}\beta_n/\sigma_\beta)]$ converges to zero. This follows from an application of Lemma B.1 in Leeb and Pötscher (2003a).

We next prove part 3. The terms in (12) involving the standard normal density ϕ are readily seen to converge to $\sigma^2_{\alpha,\infty}\rho^2_\infty r\phi(r)$ with the convention that $r\phi(r) = 0$ if $r = \pm\infty$. Furthermore, we see from Proposition A.1 that $\Phi(c - \sqrt{n}\beta_n/\sigma_\beta) - \Phi(-c - \sqrt{n}\beta_n/\sigma_\beta)$ converges to $\Phi(r)$ and that $\sigma^2_\alpha\rho^2(n(\beta_n/\sigma_\beta)^2 - 1)$ converges to $\sigma^2_{\alpha,\infty}(\chi^2 - \rho^2_\infty)$ (which may be infinite). This proves the result provided χ is finite or $\Phi(r)$ is positive. If χ is infinite and $\Phi(r)$ is zero (i.e., if $r = -\infty$), Lemma B.1 in Leeb and Pötscher (2003a) shows that the third term in (12) converges to the claimed limit.

Part 4 is proved analogously to part 3. ■

Remark A.7. In the important case where $\rho_\infty \neq 0$ the following simplifications arise: If $\rho_\infty \neq 0$ and $\zeta \neq 0$ in part 1 of the proposition, then necessarily $\chi = \text{sign}(\rho_\infty\zeta)\infty$, and hence *MSE* converges to ∞. If $\rho_\infty \neq 0$ in part 3 of the proposition, then necessarily $\chi = \text{sign}(\rho_\infty)\infty$, and hence *MSE* converges to ∞ provided $r > -\infty$. If $\rho_\infty \neq 0$ in part 4 of the proposition, then necessarily $\chi = \text{sign}(-\rho_\infty)\infty$, and hence *MSE* converges to ∞ provided $s > -\infty$.

Remark A.8. The preceding propositions in fact allow for a characterization of all possible accumulation points of the model selection probabilities, the finite-sample distribution, the (scaled) bias, and the (scaled) mean-squared error of the post-model-selection estimator under arbitrary sequences of parameters (α_n, β_n): Given any sequence (α_n, β_n), compactness of $\mathbb{R} \cup \{-\infty,\infty\}$ implies that every subsequence (n_i) contains a further subsequence $(n_{i(j)})$ such that the quantities $\sqrt{n}\beta_n/(\sigma_\beta c)$, $\sqrt{n}\beta_n(\rho/\sigma_\beta)$, $c - \sqrt{n}\beta_n/\sigma_\beta$, $c + \sqrt{n}\beta_n/\sigma_\beta$, and the expressions in the limit operators in Propositions A.4 and A.6 converge to respective limits in $\mathbb{R} \cup \{-\infty,\infty\}$ along the subsequence $(n_{i(j)})$. Applying the preceding propositions to the subsequence $(n_{i(j)})$ provides the desired characterization of all accumulation points.

https://doi.org/10.1017/S0266466605050036 Published online by Cambridge University Press

PROPOSITION A.9. *The post-model-selection estimator $\tilde{\alpha}$ is uniformly consistent for α, i.e.,*

$$\lim_{n\to\infty} \sup_{(\alpha,\beta)\in\mathbb{R}^2} P_{n,\alpha,\beta}(|\tilde{\alpha} - \alpha| > \varepsilon) = 0$$

for every $\varepsilon > 0$.

Proof. Using Chebychev's inequality we obtain

$$\begin{aligned}
&P_{n,\alpha,\beta}(|\tilde{\alpha} - \alpha| > \varepsilon) \\
&\quad = P_{n,\alpha,\beta}(|\hat{\alpha}(R) - \alpha| > \varepsilon, \hat{M} = R) + P_{n,\alpha,\beta}(|\hat{\alpha}(U) - \alpha| > \varepsilon, \hat{M} = U) \\
&\quad \le P_{n,\alpha,\beta}(|\hat{\alpha}(R) - \alpha| > \varepsilon, \hat{M} = R) + P_{n,\alpha,\beta}(|\hat{\alpha}(U) - \alpha| > \varepsilon) \\
&\quad \le \min\{P_{n,\alpha,\beta}(|\hat{\alpha}(R) - \alpha| > \varepsilon), P_{n,\alpha,\beta}(\hat{M} = R)\} + \sigma_\alpha^2/(n\varepsilon^2).
\end{aligned}$$

Because $\sigma_\alpha^2/(n\varepsilon^2)$ is independent of (α,β) and converges to zero, it suffices to show that the first term on the far r.h.s. of the preceding display converges to zero uniformly in (α,β). Observe that $\hat{\alpha}(R) - \alpha$ is distributed normally with mean $(-\rho\sigma_\alpha/\sigma_\beta)\beta$ and variance $\sigma_\alpha^2(1-\rho^2)/n$. In view of (3), the first term on the far r.h.s. of the preceding display hence equals

$$\min\{1 - \Delta(\sqrt{n}\rho(1-\rho^2)^{-1/2}\beta/\sigma_\beta, \sqrt{n}\sigma_\alpha^{-1}(1-\rho^2)^{-1/2}\varepsilon), \Delta(\sqrt{n}\beta/\sigma_\beta, c)\}, \tag{A.6}$$

which clearly does not depend on the value of the parameter α. Now

$$\lim_{n\to\infty} \sup_{|\beta|\ge 2c\sigma_\beta/\sqrt{n}} \Delta(\sqrt{n}\beta/\sigma_\beta, c) = 0$$

by an application of Proposition A.1. Furthermore,

$$\begin{aligned}
&\sup_{|\beta|<2c\sigma_\beta/\sqrt{n}} [1 - \Delta(\sqrt{n}\rho(1-\rho^2)^{-1/2}\beta/\sigma_\beta, \sqrt{n}\sigma_\alpha^{-1}(1-\rho^2)^{-1/2}\varepsilon)] \\
&\quad \le 2 - 2\Phi((1-\rho^2)^{-1/2}(-2c|\rho| + \sigma_\alpha^{-1}\sqrt{n}\varepsilon)),
\end{aligned}$$

which converges to zero because $\varepsilon > 0$, $\rho \to \rho_\infty$, and because $c/\sqrt{n} \to 0$. It now follows that (A.6) converges to zero uniformly. ■

APPENDIX B: ASYMPTOTIC RESULTS FOR CONSERVATIVE MODEL SELECTION PROCEDURES

In the following discussion we consider the linear regression model (1) under the assumptions of Section 2. Furthermore, we assume as in Section 2.2 that c does not depend on sample size and satisfies $0 < c < \infty$.

https://doi.org/10.1017/S0266466605050036 Published online by Cambridge University Press

PROPOSITION B.1. *The post-model-selection estimator $\tilde{\alpha}$ is uniformly consistent for α, i.e.,*

$$\lim_{n\to\infty} \sup_{(\alpha,\beta)\in\mathbb{R}^2} P_{n,\alpha,\beta}(|\tilde{\alpha} - \alpha| > \varepsilon) = 0$$

for every $\varepsilon > 0$.

Proof. The proof is identical to the proof of Proposition A.9 up to and including (A.6). Now

$$\lim_{\gamma\to\infty} \limsup_{n\to\infty} \sup_{|\beta|\geq\gamma\sigma_\beta/\sqrt{n}} \Delta(\sqrt{n}\beta/\sigma_\beta, c) = 0$$

as a consequence of Lemma C.3 in Leeb and Pötscher (2003b). Furthermore,

$$\sup_{|\beta|<\gamma\sigma_\beta/\sqrt{n}} [1 - \Delta(\sqrt{n}\rho(1-\rho^2)^{-1/2}\beta/\sigma_\beta, \sqrt{n}\sigma_\alpha^{-1}(1-\rho^2)^{-1/2}\varepsilon)]$$
$$\leq 2 - 2\Phi((1-\rho^2)^{-1/2}(-\gamma|\rho| + \sigma_\alpha^{-1}\sqrt{n}\varepsilon)),$$

which converges to zero for every given $\gamma \in \mathbb{R}$ because $\varepsilon > 0$ and $\rho \to \rho_\infty$. It then follows that (A.6) converges to zero uniformly. ■

APPENDIX C: THE MAXIMAL ABSOLUTE BIAS AND THE MAXIMAL MSE ARE UNBOUNDED FOR GENERAL CONSISTENT MODEL SELECTION PROCEDURES

We give here a simple proof of the fact that the (scaled) maximal absolute bias and hence the (scaled) maximal mean-squared error of a post-model-selection estimator diverges to infinity if an arbitrary consistent model selection procedure is employed. This is a variant of the result of Yang (2003), who uses a predictive mean-square risk measure instead. Our proof is based on the contiguity argument discussed in Remark 4.4. An advantage of this proof is that—contrary to Yang's proof—it does not rely on a normality assumption for the errors.

We assume the simple linear regression model (1) under the basic assumptions made in Section 2, except that the errors ϵ_t only need to be i.i.d. with mean zero and (finite) variance $\sigma^2 > 0$. (The assumption that σ^2 is known is inessential here. If σ^2 is unknown, and hence f depends on the scale parameter σ, Proposition C.1 holds for every value of σ^2.) Furthermore, we assume that ϵ_t has a density f that possesses an absolutely continuous derivative f' satisfying

$$0 < \int_{-\infty}^{\infty} (f'(x)/f(x))^2 f(x)\, dx < \infty.$$

https://doi.org/10.1017/S0266466605050036 Published online by Cambridge University Press

Note that the conditions on f guarantee that the information of f is finite and positive. (These conditions are obviously satisfied in the special case of normally distributed errors.) Let $\check{M}$ now be an arbitrary model selection procedure that consistently selects between the models R and U. Furthermore, let $\check{\alpha}$ denote the corresponding post-model-selection estimator (i.e., $\check{\alpha} = \hat{\alpha}(R)$ if $\check{M} = R$ and $\check{\alpha} = \hat{\alpha}(U)$ if $\check{M} = U$). In the following $E_{n,\alpha,\beta}$ denotes the expectation operator w.r.t. $P_{n,\alpha,\beta}$. Recall that ρ_∞ is less than unity in absolute value because the limit Q of $X'X/n$ has been assumed to be positive definite.

PROPOSITION C.1. *Suppose that $\rho_\infty \neq 0$. Then the maximal absolute bias $\sup_{\alpha,\beta}|E_{n,\alpha,\beta}[\sqrt{n}(\check{\alpha} - \alpha)]|$, and hence the maximal mean-squared error $\sup_{\alpha,\beta} E_{n,\alpha,\beta}[n(\check{\alpha} - \alpha)^2]$, goes to infinity for $n \to \infty$.*

Proof. Clearly, it suffices to prove the result for the maximal absolute bias. The following elementary relations hold:

$$\begin{aligned}
&E_{n,\alpha,\beta}[\sqrt{n}(\check{\alpha} - \alpha)] \\
&\quad = E_{n,\alpha,\beta}[\sqrt{n}(\hat{\alpha}(R) - \alpha)\mathbf{1}(\check{M} = R)] + E_{n,\alpha,\beta}[\sqrt{n}(\hat{\alpha}(U) - \alpha)\mathbf{1}(\check{M} = U)] \\
&\quad = E_{n,\alpha,\beta}[\sqrt{n}(\hat{\alpha}(R) - \alpha)] + E_{n,\alpha,\beta}[\sqrt{n}(\hat{\alpha}(U) - \hat{\alpha}(R))\mathbf{1}(\check{M} = U)] \\
&\quad = E_{n,\alpha,\beta}[\sqrt{n}(\hat{\alpha}(R) - \alpha)] + \rho(\sigma_\alpha/\sigma_\beta)E_{n,\alpha,\beta}[\sqrt{n}\hat{\beta}(U)\mathbf{1}(\check{M} = U)].
\end{aligned}$$

Furthermore,

$$E_{n,\alpha,\beta}[\sqrt{n}(\hat{\alpha}(R) - \alpha)] = \sqrt{n}\beta \sum_{t=1}^{n} x_{t1}x_{t2} \Big/ \sum_{t=1}^{n} x_{t1}^2 = -\sqrt{n}\beta\rho\sigma_\alpha\sigma_\beta^{-1}.$$

Consequently, for every α and every $r \in \mathbb{R}$ we have

$$\begin{aligned}
&\liminf_{n\to\infty} \sup_{\beta\in\mathbb{R}} |E_{n,\alpha,\beta}[\sqrt{n}(\check{\alpha} - \alpha)]| \\
&\quad \geq \liminf_{n\to\infty} |E_{n,\alpha,r/\sqrt{n}}[\sqrt{n}(\check{\alpha} - \alpha)]| \\
&\quad \geq \liminf_{n\to\infty} |E_{n,\alpha,r/\sqrt{n}}[\sqrt{n}(\hat{\alpha}(R) - \alpha)]| = |r||\rho_\infty|\sigma_{\alpha,\infty}\sigma_{\beta,\infty}^{-1}, \qquad \textbf{(C.1)}
\end{aligned}$$

provided we can show that

$$\limsup_{n\to\infty} |E_{n,\alpha,r/\sqrt{n}}[\sqrt{n}(\hat{\beta}(U))\mathbf{1}(\check{M} = U)]| = 0 \qquad \textbf{(C.2)}$$

for every $r \in \mathbb{R}$. We apply the Cauchy–Schwartz inequality to obtain

$$|E_{n,\alpha,r/\sqrt{n}}[\sqrt{n}(\hat{\beta}(U))\mathbf{1}(\check{M} = U)]| \leq E_{n,\alpha,r/\sqrt{n}}^{1/2}[n(\hat{\beta}(U))^2][P_{n,\alpha,r/\sqrt{n}}(\check{M} = U)]^{1/2}. \qquad \textbf{(C.3)}$$

The first term on the r.h.s. in (C.3) is easily seen to satisfy

$$E_{n,\alpha,r/\sqrt{n}}^{1/2}[n(\hat{\beta}(U))^2] = (\sigma_\beta^2 + r^2)^{1/2}.$$

https://doi.org/10.1017/S0266466605050036 Published online by Cambridge University Press

To prove (C.2) it hence suffices to show that $\limsup_{n\to\infty} P_{n,\alpha,r/\sqrt{n}}(\check{M} = U) = 0$. Because the model is locally asymptotically normal (Koul and Wang, 1984, Theorem 2.1 and Remark 1; Hajek and Sidak, 1967, p. 213), the sequence of probability measures $P_{n,\alpha,r/\sqrt{n}}$ is contiguous w.r.t. the sequence $P_{n,\alpha,0}$ (for every $r \in \mathbb{R}$). Because $\limsup_{n\to\infty} P_{n,\alpha,0}(\check{M} = U) = 0$ by the assumed consistency of the model selection procedure, contiguity implies

$$\limsup_{n\to\infty} P_{n,\alpha,r/\sqrt{n}}(\check{M} = U) = 0$$

for every $r \in \mathbb{R}$, cf. Remark 4.4. This establishes (C.2) and hence (C.1). Letting $|r|$ go to infinity in (C.1) then completes the proof (note that $|\rho_\infty|$ and $\sigma_{\alpha,\infty}$ are positive and $\sigma_{\beta,\infty}^{-1}$ is finite). ■

Remark C.2.

1. The proof in fact shows that this result holds for fixed α and any bounded neighborhood of $\beta = 0$, i.e., $\sup_{|\beta|\le s} |E_{n,\alpha,\beta}[\sqrt{n}(\check{\alpha} - \alpha)]|$ and $\sup_{|\beta|\le s} E_{n,\alpha,\beta}[n(\check{\alpha} - \alpha)^2]$ diverge to infinity as $n \to \infty$ for each fixed α and $s > 0$.
2. The preceding proposition is formulated for the simple regression model with two regressors and only two competing models from which to choose. It can easily be extended to more general cases. The preceding proof should also easily extend to the risk measure used in Yang (2003). We do not pursue these issues here.

https://doi.org/10.1017/S0266466605050036 Published online by Cambridge University Press