Discussion

Hemant ISHWARAN and J. Sunil RAO

In reading the two articles written by Hjort and Claeskens, readers will find several important and fundamental topics in statistics discussed, including model selection, inference after model selection (post–model selection), and the notion of model averaging (from both frequentist and Bayesian perspectives). Without a doubt the authors have selected a wide range of ambitious topics to study in their work, and ones, we should add, that are very likely to create a lively set of discussions! We thank the editor, Francisco Samaniego, and the other members of the editorial staff for giving us the opportunity to comment on these interesting papers.

At the heart of Claeskens and Hjort's approach is the clever idea of a *local asymptotic misspecification framework*. The authors ask the following question: If we have a parametric model involving a parameter of interest $\boldsymbol{\theta} \in \Re^p$ and a "nuisance" parameter $\boldsymbol{\gamma} \in \Re^q$, what is the effect on inference for $\boldsymbol{\theta}$ when $\boldsymbol{\gamma}$ is subjected to some kind of selection procedure? To study this question, the authors look at the asymptotic limiting distribution for restricted maximum likelihood estimators (or model-averaged estimators) under a changing sequence of alternative models for $\boldsymbol{\gamma}$ (the local misspecification framework). These limits are then used to determine asymptotic mean squared error performance, which can then be used to decide between estimators. Lower mean squared error performance translates into a form of robustness to misspecification. This whole approach is not restricted to just estimating $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ but is more generally discussed in terms of some functional $\mu(\boldsymbol{\theta}, \boldsymbol{\gamma})$, the so-called *focused parameter.*

1. DOES THE LOCAL MISSPECIFICATION FRAMEWORK REALLY ADDRESS VARIABLE SELECTION PROBLEMS?

One way to view the local asymptotic framework is that it is a method for mimicking the effects of model uncertainty. Model uncertainty here represents the excess variance incurred when choosing among a set of parameters whose values are not known a priori. The local asymptotics framework is certainly well poised to address this issue. However, although we feel that the method is well motivated in most examples considered in the two articles, our concern is that the method is not being appropriately applied in the examples concerned with variable selection. As we will argue, these examples only look at settings when all the true nonzero regression parameters have been included in the model and when model selection is restricted to the zero coefficients. To us this is not a realistic subset selection problem.

A concrete example will illustrate our point. Suppose we have the usual linear regression setup where we are given n independent responses Y_i, with corresponding K-dimensional covariates $\mathbf{x}_i$. The problem is to find the subset of covariate parameters from $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)^t$ that are nonzero where it is assumed that

$$Y_i = \mathbf{x}_i^t \boldsymbol{\beta} + \epsilon_i, \qquad i = 1, \ldots, n, \tag{1}$$

and ϵ_i are independent random variables such that $\mathbb{E}(\epsilon_i) = 0$ and $\mathbb{E}(\epsilon_i) = \sigma^2$. For our later discussion we will not need to make any distributional assumptions for ϵ_i, but for the moment let us suppose that ϵ_i are normally distributed. Then, due to the assumption of normality, the maximum likelihood estimator is the OLS (ordinary least squares) estimator. Thus, restricted maximum likelihood estimation corresponds to restricted OLS. That is, the general submodel estimator $\hat{\mu}_S = \mu(\hat{\boldsymbol{\theta}}_S, \hat{\boldsymbol{\gamma}}_S, \boldsymbol{\gamma}_{0,S^c})$ will depend on the restricted OLS estimator for $\boldsymbol{\beta}$. Consequently, to understand how the theory applies, we can consider how it applies to the OLS and restricted OLS.

Let us identify what $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ and $(\boldsymbol{\theta}_0, \boldsymbol{\gamma}_0)$ are in this problem and what the misspecification framework is (we will ignore the parameter σ^2 as it plays a limited role in this argument). First notice that the true value $\boldsymbol{\gamma}_0$ of $\boldsymbol{\gamma}$ must be some fixed known value (otherwise one could not compute $\hat{\mu}_S$, as this requires knowing the value $\boldsymbol{\gamma}_{0,S^c}$). A nonzero value would be associated with some kind of offset value, which is of limited interest in a variable selection setting, which leaves the only other possibility, $\boldsymbol{\gamma}_0 = \mathbf{0}_q$, where $\mathbf{0}_q$ is the q-dimensional vector whose coordinates all equal 0 (we will show shortly that there is another important reason $\boldsymbol{\gamma}_0$ must be zero). So because $\boldsymbol{\gamma}_0 = \mathbf{0}_q$, this means that $\boldsymbol{\theta}_0$, the true value for $\boldsymbol{\theta}$, contains all the nonzero coefficients of the model (and possibly some that are 0). As the misspecification framework looks at models whose $\boldsymbol{\gamma}$ parameters are perturbed around $\boldsymbol{\gamma}_0$, this means that we are interested in the asymptotics of the restricted OLS estimator under misspecified models of the form

$$f_{i,\text{true}}(y|\mathbf{x}_i) = \phi(Y_i|\mathbf{x}_i^t \boldsymbol{\beta}_n, \sigma^2),$$

where $\phi(\cdot|m, \sigma^2)$ denotes a normal density with mean m and variance σ^2 and

$$\boldsymbol{\beta}_n = (\boldsymbol{\theta}_n^t, \boldsymbol{\gamma}_n^t)^t = (\boldsymbol{\theta}_0^t, \boldsymbol{\delta}^t/\sqrt{n})^t, \quad \text{where } \boldsymbol{\delta} \in \Re^q.$$

Observe that because we are perturbing $\boldsymbol{\gamma}$ around $\boldsymbol{\gamma}_0 = \mathbf{0}_q$ and because $\boldsymbol{\theta}_0$ contains all the nonzero coefficients of $\boldsymbol{\beta}$, the misspecification framework implies we are studying the effect of model uncertainty when *we have specified a model that includes all the nonzero coefficients in the model and subset selection is over the zero coefficients.* This seems unnatural for the following reasons: (a) Why should subset selection be restricted to coefficients known to be 0? (b) It requires that one is lucky enough to have not underestimated the model.

Hemant Ishwaran is Associate Staff, Department of Biostatistics and Epidemiology Wb4, Cleveland Clinic Foundation, Cleveland, OH 44195 (E-mail: *ishwaran@bio.ri.ccf.org*). J. Sunil Rao is Associate Professor, Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH 44106 (E-mail: *sunil@hal.epbi.cwru.edu*).

Journal of the American Statistical Association
December 2003, Vol. 98, No. 464, Theory and Methods
DOI 10.1198/016214503000000855

2. THE ASYMPTOTICS CAN BREAK DOWN

Why does the misspecification framework require that $\boldsymbol\gamma_0=\mathbf{0}_q$? The answer turns out to be quite simple. If one misspecifies the model by excluding a nonzero coefficient of $\boldsymbol\beta$, then one cannot obtain a proper limiting distribution in general, and thus the whole idea of looking at the limiting distribution to quantify the effect of model uncertainty will fail. Note that these problems are not at all specific to the linear regression setup. They apply to generalized linear models, such as the Poisson and logistic regression models illustrated in the articles, for all the same reasons.

Here is a more formal way to see this in the context of our linear regression example. Let $\boldsymbol\beta=(\boldsymbol\theta_0^t,\boldsymbol\gamma_0^t)^t$. To keep the notation simple, we will consider only nested subsets S of the form

$$\varnothing,\{1\},\{1,2\},\ldots,\{1,2,\ldots,q\}.$$

By the notation $S=\{1,2,\ldots,l\}$ we mean that the restricted estimator $\hat{\boldsymbol\beta}_S$ is computed from the restricted OLS based on the first $k=p+l$ coefficients, where $l=0,1,\ldots,q$ (the null set $\varnothing$ corresponds to $l=0$ and $k=p$). More formally, rewrite (1) as $\mathbf{Y}=\mathbf{X}\boldsymbol\beta+\boldsymbol\epsilon$, where $\mathbf{X}$ is the $n\times K$ design matrix ($K=p+q$) and $\mathbf{Y}=(Y_1,\ldots,Y_n)^t$ and $\boldsymbol\epsilon=(\epsilon_1,\ldots,\epsilon_n)^t$. Then the restricted estimator based on $S=\{1,2,\ldots,l\}$ is

$$\hat{\boldsymbol\beta}_S=\big(\hat{\boldsymbol\beta}[k]^t,\mathbf{0}_{q-l}^t\big)^t, \tag{2}$$

where $\hat{\boldsymbol\beta}[k]=(\mathbf{X}[k]^t\mathbf{X}[k])^{-1}\mathbf{X}[k]^t\mathbf{Y}$ and $\mathbf{X}[k]$ is the $n\times k$ matrix composed of the first k columns of $\mathbf{X}$. The estimator (2) assumes that $\boldsymbol\gamma_0=\mathbf{0}_q$. In general, if $\boldsymbol\gamma_0$ is nonzero, the restricted estimator is

$$\hat{\boldsymbol\beta}_S=\big(\hat{\boldsymbol\beta}[k]^t,\boldsymbol\gamma_0[-l]^t\big)^t,$$

where $\boldsymbol\gamma_0[-l]=(\gamma_{0,l+1},\ldots,\gamma_{0,q})^t$.

Now we show why it is necessary to assume that $\boldsymbol\gamma_0$ is zero. Let $\boldsymbol\beta[k]=(\beta_1,\ldots,\beta_k)^t$ and $\boldsymbol\beta[-k]=(\beta_{k+1},\ldots,\beta_K)^t$. Some simple algebra shows that

$$\hat{\boldsymbol\beta}[k]=\boldsymbol\beta[k]+(\mathbf{X}[k]^t\mathbf{X}[k])^{-1}\mathbf{X}[k]^t\mathbf{X}[-k]\boldsymbol\beta[-k] + (\mathbf{X}[k]^t\mathbf{X}[k])^{-1}\mathbf{X}[k]^t\boldsymbol\epsilon,$$

where $\mathbf{X}[-k]$ refers to the $n\times(K-k)$ matrix formed by excluding the first k columns of $\mathbf{X}$. So far we have assumed that ϵ_i are normally distributed. However, the following argument will hold without this condition. Hereafter we will assume only that ϵ_i are independent random variables such that $\mathbb{E}(\epsilon_i)=0$, $\mathbb{E}(\epsilon_i)=\sigma^2$, and $\mathbb{E}(\epsilon_i^4)\le M$ for some $M<\infty$. We also need the following mild conditions for the covariates:

$$\max_{1\le i\le n}\|\mathbf{x}_i\|/\sqrt{n}\to 0 \quad\text{and}\quad \mathbf{X}^t\mathbf{X}/n\to\mathbf{Q},$$

where $\mathbf{Q}$ is positive definite and $\|\cdot\|$ denotes the ℓ_2 norm. Under these conditions one can show that

$$\hat{\boldsymbol\beta}[k]=\boldsymbol\beta[k]+\boxed{\mathbf{Q}^{-1}[k:k]\mathbf{Q}[k:-k]\boldsymbol\beta[-k]} + o(1)+O_p(n^{-1/2}), \tag{3}$$

where $\mathbf{Q}$ has been partitioned according to

$$\mathbf{Q}=\begin{pmatrix}\mathbf{Q}[k:k] & \mathbf{Q}[k:-k]\\ \mathbf{Q}[-k:k] & \mathbf{Q}[-k:-k]\end{pmatrix}.$$

(For example, $\mathbf{Q}[k:k]$ is the upper left $k\times k$ submatrix of $\mathbf{Q}$.) The second term in (3), highlighted by the rectangular box, is the culprit. It represents a bias term that in general does not vanish. In fact, because $\mathbf{Q}$ is positive definite this term can be zero only if $\mathbf{Q}=\mathbf{I}$ or if $\mathbf{Q}[k:-k]\boldsymbol\beta[-k]=\mathbf{0}_k$. Consider the misspecified model associated with $\boldsymbol\beta_n=(\boldsymbol\theta_n^t,\boldsymbol\gamma_n^t)^t$, where $\boldsymbol\gamma_n=\boldsymbol\gamma_0+\boldsymbol\delta/\sqrt{n}$. Suppose that $S=\{1,2,\ldots,l\}$. Then, unless $\mathbf{Q}=\mathbf{I}$ or $\mathbf{Q}[k:-k]\boldsymbol\beta_n[-k]=O(1/\sqrt{n})$,

$$\|\sqrt{n}(\hat{\boldsymbol\beta}_S-\boldsymbol\beta_n)\|\ge\|\sqrt{n}(\hat{\boldsymbol\beta}[k]-\boldsymbol\beta_n[k])\|\xrightarrow{p}\infty.$$

Thus, the asymptotics will break down. Because $\boldsymbol\beta_n[-k]=\boldsymbol\gamma_n[-l]$ one way to avoid this problem is to assume that $\boldsymbol\gamma_n[-l]=O(1/\sqrt{n})$, which implies that $\boldsymbol\gamma_0[-l]=\mathbf{0}_{q-l}$. Because l is arbitrary this implies that $\boldsymbol\gamma_0=\mathbf{0}_q$.

Remark 1. Our argument shows this problem exists even if $\boldsymbol\theta$ is perturbed as suggested in Remark 4.1 of the "Frequentist Model Average Estimators" article. Moreover, the same problems apply to the model-averaged estimators discussed in both articles.

3. WHAT IS BETTER: FORWARD OR BACKWARD STEPWISE REGRESSION?

The previous discussion does indicate, however, that the local asymptotics framework may be applicable in the orthogonal linear regression setup without resorting to the assumption that $\boldsymbol\gamma_0=\mathbf{0}_q$. Perhaps the authors could comment on this point? We will also consider the orthogonal case; however, we will take a different approach by using a method introduced by Pötscher (1991). This method is quite different from the local asymptotics setup. Rather than looking at well-behaved $\sqrt{n}$-asymptotic distributions, the idea is to consider the effects of model selection under a procedure that is *inconsistent.* This has the advantage that it will allow us to study focus parameters μ that do not converge at a $\sqrt{n}$ rate. It will also lend some insight into a question we have always wondered about: namely, is it better to use forward or backward stepwise regression? We will give a brief outline of the argument in the context of nested subset selection. For technical details, proofs, and a more extended discussion, see Ishwaran and Rao (2003).

For this result we assume that $\mathbf{X}^t\mathbf{X}/n=\mathbf{I}$ and that the coordinates of $\boldsymbol\beta$ have been ordered so that the first k_0 coordinates are the nonzero values. That is,

$$\boldsymbol\beta=(\beta_1,\ldots,\beta_{k_0},\mathbf{0}_{K-k_0}^t)^t.$$

However, unlike the previous setup, the value for k_0, the complexity of the model, is assumed to be unknown (our only assumption being that $1\le k_0\le K$; previously it was assumed that $k_0\le p$ where p was known). The complexity k_0 will be our focus parameter μ. Observe that μ is nondifferentiable.

Pötscher (1991), and more recently Leeb and Pötscher (2003), studied the effects of selection bias from a backward stepwise procedure. Let $\hat{\boldsymbol\beta}=(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}$ be the (unrestricted) OLS estimator of $\boldsymbol\beta$ and let $\hat\sigma_n^2=\|\mathbf{Y}-\mathbf{X}\hat{\boldsymbol\beta}\|^2/(n-K)$ be the unbiased estimator for σ^2 based on the full model. To test whether $\hat\beta_k$, the kth coefficient of the OLS, is 0, define the following test statistic:

$$Z_{k,n}=\frac{\sqrt{n}\hat\beta_k}{\hat\sigma_n}.$$

Let $\alpha_1, \dots, \alpha_K$ be a sequence of fixed positive α-significance values for the $Z_{k,n}$ test statistics. Let $z_{\alpha/2}$ be the 100× $(1-\alpha/2)$ percentile of a standard normal distribution. Estimate the true complexity k_0 by the estimator $\hat{k}_B$, where

$$\hat{k}_B = \max\{k : |Z_{k,n}| \ge z_{\alpha_k/2}, k = 0, \dots, K\},$$

and where, to ensure that $\hat{k}_B$ is well defined, we take $Z_{0,n} = 0$ and $z_{\alpha_0/2} = 0$. Observe that if $\hat{k}_B = k$, then $Z_{k,n}$ is the first test statistic starting from $k = K$ and going to $k = 0$ such that $|Z_{k,n}| \ge z_{\alpha_k/2}$ and $|Z_{j,n}| < z_{\alpha_j/2}$ for $j = k+1, \dots, K$. This corresponds to accepting the event $\{\boldsymbol{\beta} : \beta_{k+1} = 0, \dots, \beta_K = 0\}$ but rejecting $\{\boldsymbol{\beta} : \beta_k = 0, \dots, \beta_K = 0\}$. The post-model selection estimator for $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}_B = \mathbf{0}_K \mathbb{I}\{\hat{k}_B = 0\} + \sum_{k=1}^{K} \left(\hat{\boldsymbol{\beta}}[k]^t, \mathbf{0}_{K-k}^t\right)^t \mathbb{I}\{\hat{k}_B = k\}, \tag{4}$$

where $\mathbb{I}(\cdot)$ is the indicator function. It should be clear that $\hat{k}_B$ and $\hat{\boldsymbol{\beta}}_B$ are derived from a backward stepwise mechanism.

A forward stepwise procedure can be defined in an analogous way. Define

$$\hat{k}_F = \min\{k - 1 : |Z_{k,n}| < z_{\alpha_k/2}, k = 1, \dots, K+1\},$$

where $Z_{K+1,n} = 0$ and $\alpha_{K+1} = 0$ are chosen to ensure a well-defined procedure. Observe that if $\hat{k}_F = k - 1$, then $Z_{k,n}$ is the first test statistic such that $|Z_{k,n}| < z_{\alpha_k/2}$ and $|Z_{j,n}| \ge z_{\alpha_j/2}$ for $j = 1, \dots, k-1$. This corresponds to accepting the event $\{\boldsymbol{\beta} : \beta_1 \ne 0, \dots, \beta_{k-1} \ne 0\}$ but rejecting $\{\boldsymbol{\beta} : \beta_1 \ne 0, \dots, \beta_k \ne 0\}$. Note that $\hat{k}_F = 0$ if $|Z_{1,n}| < z_{\alpha_1/2}$. The post-model selection estimator $\hat{\boldsymbol{\beta}}_F$ derived from $\hat{k}_F$ is defined analogously to (4).

We now state some asymptotic properties of $\hat{k}_B$ and $\hat{k}_F$. Part (a) of the following theorem is related to Lemma 4 of Pötscher (1991).

Theorem 1 (Ishwaran and Rao, 2003). Assume that $\mathbf{X}^t\mathbf{X}/n = \mathbf{I}$ and $\max_{1 \le i \le n} \|\mathbf{x}_i\|/\sqrt{n} \to 0$. Also assume that ϵ_i are independent such that $\mathbb{E}(\epsilon_i) = 0$, $\mathbb{E}(\epsilon_i^2) = \sigma^2$, and $\mathbb{E}(\epsilon_i^4) \le M$ for some $M < \infty$. Let k_B and k_F denote the limits for $\hat{k}_B$ and $\hat{k}_F$, respectively, as $n \to \infty$. For $1 \le k \le K$,

$$\text{(a)} \quad \mathbb{P}\{k_B = k\} = 0 \times \mathbb{I}\{k < k_0\} + (1 - \alpha_{k_0+1}) \cdots (1 - \alpha_K)\mathbb{I}\{k = k_0\} + \alpha_k(1 - \alpha_{k+1}) \cdots (1 - \alpha_K)\mathbb{I}\{k > k_0\},$$

$$\text{(b)} \quad \mathbb{P}\{k_F = k\} = 0 \times \mathbb{I}\{k < k_0\} + (1 - \alpha_{k_0+1})\mathbb{I}\{k = k_0\} + (1 - \alpha_{k+1})\alpha_{k_0+1} \cdots \alpha_k \mathbb{I}\{k > k_0\},$$

where $\alpha_{K+1} = 0$ in (b).

Theorem 1 can be used to assess the performance of the two procedures. Suppose that $\alpha_k = \alpha > 0$ for each k. Then the limiting probability of correctly recovering the true complexity is $\mathbb{P}\{k_F = k_0\} = (1 - \alpha)$ for forward stepwise, whereas $\mathbb{P}\{k_B = k_0\} = (1 - \alpha)^{K-k_0}$ for backward stepwise. Notice if $K - k_0$ is large, this last probability can be approximated by $\exp(-(K - k_0)\alpha)$, which becomes exponentially small as K increases. Simply put, backward stepwise can lead to models that are much too large. To see visually the effect of model uncertainty, consider Figure 1, which presents the limiting probabilities for the two procedures under various choices of K and k_0 (all figures computed with $\alpha = .10$). One can clearly see how much better the forward procedure is, especially as K becomes larger.

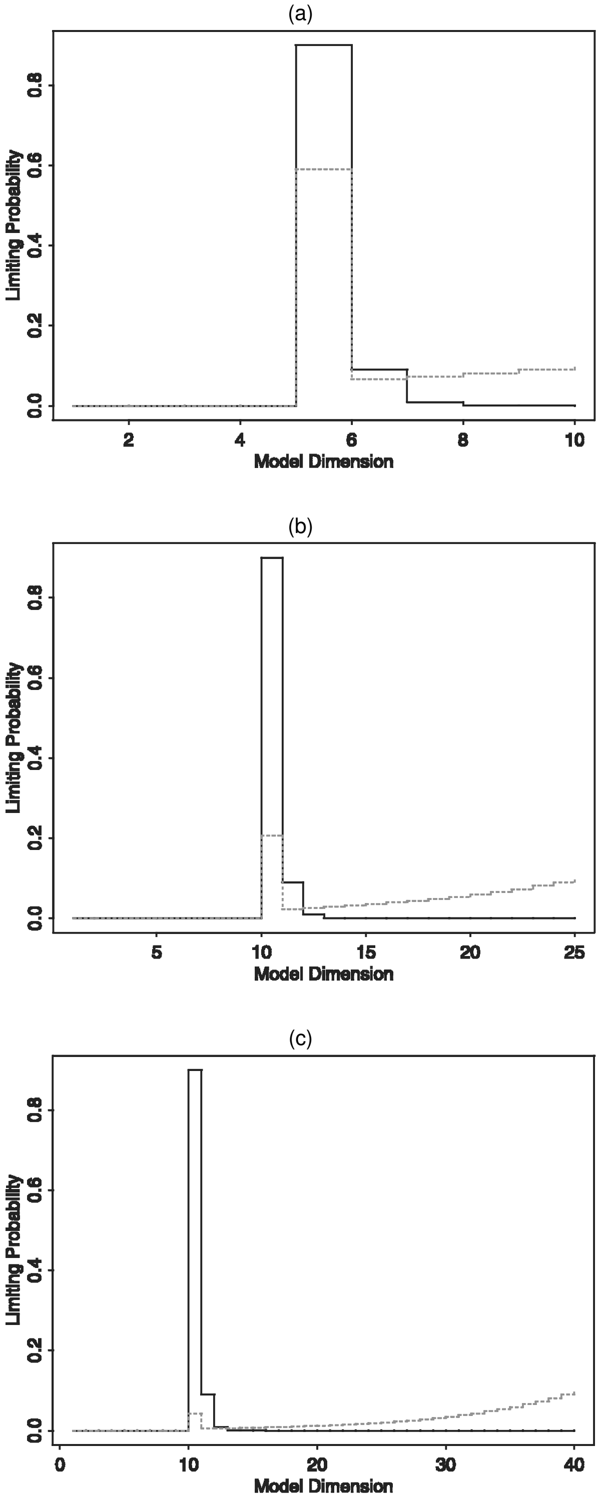

Figure 1. Effect of Model Uncertainty Under Forward and Backward Stepwise Regression. Limiting probabilities versus model dimension k for $\hat{k}_F$ (—) and $\hat{k}_B$ (· · ·). In all cases $\alpha_k = .10$. From top to bottom: (a) $K = 10$, $k_0 = 5$; (b) $K = 25$, $k_0 = 10$; (c) $K = 40$, $k_0 = 10$.

Because of limited space we only mention the effects that model uncertainty has on the performance of complexity estimators, but it should be reasonably clear from Theorem 1 that the post-model selection estimators $\hat{\boldsymbol{\beta}}_B$ and $\hat{\boldsymbol{\beta}}_F$ will have limiting mixture distributions. A more detailed analysis can be then be used to reveal the effect of selection bias due to model selection (Ishwaran and Rao 2003). For more on selection bias, see Zhang (1992) and Leeb and Pötscher (2003).

4. CONCLUSIONS

Our major concern in this discussion has been that the local misspecification framework does not correctly capture the essence of a true regression subset selection problem. This, however, should not be interpreted as a criticism of the general technique. Indeed, we are quite positive about local asymptotic arguments, seeing them as a versatile theoretical tool for studying the effects of model uncertainty. Hjort and Claeskens have illustrated one useful way this technique can be used, but we believe there are potentially many others. At least one other example we are aware of was given by Bühlmann and Yu (2002), who used a type of local misspecification framework to study bootstrap aggregation (or bagging) algorithms and their ability to sometimes reduce prediction error. This approach differs slightly because it looks at prediction error rather than a focused parameter as the main criterion of interest, but there are still interesting similarities. Bühlmann and Yu's approach was to look at the distribution of an estimator (what they called a predictor) when perturbed around a fixed value. They then derived the limiting mean squared error as the effect of the perturbation vanishes at an $O(1/\sqrt{n})$ rate. In some cases they showed that the corresponding risk is lower under a bagged version of the estimator (a type of model-averaged estimator), thus showing that model averaging can sometimes improve performance due to its ability to handle model uncertainty (there called instability).

We hope that the other discussants will indicate more examples where local asymptotics has been used. Clearly, this is an interesting technique, and one that we expect will be explored more in the future.

ADDITIONAL REFERENCES

Bühlmann, P., and Yu, B. (2002), "Understanding Bagging," *The Annals of Statistics*, 30, 927–961.

Ishwaran, H., and Rao, J. S. (2003), "Spike and Slab Variable Selection: Frequentist and Bayesian Strategies," unpublished manuscript.

Leeb, H., and Pötscher, B. M. (2003), "The Finite-Sample Distribution of Post-Model-Selection Estimators, and Uniform Versus Non-Uniform Approximations," *Econometric Theory*, 19, 100–142.

Pötscher, B. M. (1991), "Effects of Model Selection on Inference," *Econometric Theory*, 7, 163–185.

Zhang, P. (1992), "Inference After Variable Selection in Linear Regression Models," *Biometrika*, 79, 741–746.

Discussion

R. Dennis Cook and Lexin Li

1. INTRODUCTION

Claeskens and Hjort in their article "The Focused Information Criterion" (hereinafter CH) have put forth an interesting thesis, one that apparently breaks away from the sometimes confining methodology of the standard model selection paradigms. Their articles work well on a number of levels: new practically relevant ideas; fresh interpretations of standard methodology, including model averaging; and a focus that invites reflection on the foundations of model/variable selection. Their work will surely be the subject of much application and elaboration in the future. We address the focused information criterion in the following remarks.

To buy the Claeskens–Hjort focused paradigm, we must evidently be comfortable with a number of ingredients, including the following.

Known True Model. Similar to most model selection methods, the focused paradigm starts with a true model f_{true} that is known up to a finite-dimensional parameter. In many analyses f_{true} will be unknown and must be built using the observed data and prior information before addressing model selection. The process of building f_{true} can be complex and difficult to characterize, as CH mentioned in their Introduction. The focused approach seems fully applicable to the degree that the model building process can be parameterized finitely; otherwise, there may still be a substantial element of faith underlying application of the focused criterion, just as there may be for current model selection methodology. Assessing the quality of the final model is even more elusive when the methodology interweaves model building and model selection.

Nesting. The focused criterion, like much current methodology, requires that all potential models be nested within the true model. This could be an issue, depending on the application. Consider the true normal linear model with mean function $\mathbb{E}(y|x,u) = \beta_0 + \beta_1 x + (\delta/\sqrt{n})u$ and constant variance function $\text{Var}(y|x,u) = \sigma^2$. The notation here follows CH's Section 4.2, so the intercept and x are protected and $\gamma_0 = 0$. With $\mu = \mathbb{E}(y|x_0,u_0)$, we can now apply the CH machinery to contrast the full model with $\mathbb{E}(y|x,u) = \beta_0 + \beta_1 x$ and $\text{Var}(y|x,u) = \sigma^2$. However, this is not the only approach to variable selection. If we wish to understand what happens in the absence of u, then perhaps we should compare the full model to the derived submodel that conditions only on x: $\mathbb{E}(y|x) = \beta_0 +$

R. Dennis Cook is Professor, School of Statistics, University of Minnesota, St. Paul, MN 55108 (E-mail: *dennis@stat.umn.edu*). Lexin Li is Postdoctoral Fellow, Medical School, University of California, Davis, CA 95616. Cook was supported in part by National Science Foundation grant DMS 0103983.

Journal of the American Statistical Association
December 2003, Vol. 98, No. 464, Theory and Methods
DOI 10.1198/016214503000000864