Journal of Business & Economic Statistics

ISSN: 0735-0015 (Print) 1537-2707 (Online) Journal homepage: http://www.tandfonline.com/loi/ubes20

Focused Information Criteria, Model Selection, and Model Averaging in a Tobit Model With a Nonzero Threshold

Xinyu Zhang , Alan T. K. Wan & Sherry Z. Zhou

To cite this article: Xinyu Zhang , Alan T. K. Wan & Sherry Z. Zhou (2012) Focused Information Criteria, Model Selection, and Model Averaging in a Tobit Model With a Nonzero Threshold, Journal of Business & Economic Statistics, 30:1, 132-142, DOI: 10.1198/jbes.2011.10075

To link to this article: http://dx.doi.org/10.1198/jbes.2011.10075

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JBES

Focused Information Criteria, Model Selection, and Model Averaging in a Tobit Model With a Nonzero Threshold

Xinyu Zhang
Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China
(*xinyu@amss.ac.cn*)

Alan T. K. Wan and Sherry Z. Zhou
Department of Management Sciences, City University of Hong Kong, Kowloon, Hong Kong
(*msawan@cityu.edu.hk*; *zhefzhou@cityu.edu.hk*)

Claeskens and Hjort (2003) have developed a focused information criterion (FIC) for model selection that selects different models based on different focused functions with those functions tailored to the parameters singled out for interest. Hjort and Claeskens (2003) also have presented model averaging as an alternative to model selection, and suggested a local misspecification framework for studying the limiting distributions and asymptotic risk properties of post-model selection and model average estimators in parametric models. Despite the burgeoning literature on Tobit models, little work has been done on model selection explicitly in the Tobit context. In this article we propose FICs for variable selection allowing for such measures as mean absolute deviation, mean squared error, and expected expected linear exponential errors in a type I Tobit model with an unknown threshold. We also develop a model average Tobit estimator using values of a smoothed version of the FIC as weights. We study the finite-sample performance of model selection and model average estimators resulting from various FICs via a Monte Carlo experiment, and demonstrate the possibility of using a model screening procedure before combining the models. Finally, we present an example from a well-known study on married women's working hours to illustrate the estimation methods discussed. This article has supplementary material online.

KEY WORDS: Backward elimination; Censored regression; LINEX errors; Local misspecification; Mean absolute deviation; Mean squared error; Model screening; Monte Carlo.

1. INTRODUCTION

The Tobit model is now a standard approach to model-censored dependent variables. There have been many extensions of the original Tobit model based on the seminal contribution of Tobin (1958), and numerous applications of these models have appeared in economics since the 1970s. [See Amemiya (1984, 1985) for a survey of the early landmark studies.] But despite the extensive literature on Tobit modeling, few authors have explicitly examined model selection in the Tobit setting. An exception is Kao (1986), who showed that the variable selection problem in a type I Tobit model may be converted into a linear regression model using an EM algorithm. Some goodness-of-fit measures for Tobit models also have been developed by Amemiya (1981), Heckman (1984), Dhrymes (1986), and Laitila (1993). In practice, the most widely accepted model selection procedures for Tobit models are those in the information criteria framework, such as the Akaike information criterion (AIC) and Schwartz's Bayes information criterion (BIC). These methods have been well studied and arguably are the most actively used model selection strategies in all areas of statistics. The AIC and BIC each have specific strengths and weaknesses (Yang 2001; Yuan and Yang 2005; Claeskens and Hjort 2008a, 2008b). That being said, in a typical modeling situation, the "best" model chosen by the AIC or BIC is used to explain all aspects of the mechanisms underlying the data (Claeskens and Hjort 2008a, p. 145). On the other hand, it is not uncommon to find that a model good for estimating one estimand may be bad for another; Claeskens and Hjort (2003) and Claeskens, Croux, and Venkerckhoven (2006) gave some specific examples in biostatistics in which no one single model is good for every patient subgroup.

A recent article by Claeskens and Hjort (2003) took the view that a model selection method should be tailored to the parameters singled out for interest; that is, the method should select the model best suited for a given focus. Other authors, including Hand and Vinciotti (2003), Hansen (2005), and Vaida and Blanchard (2005), also have expressed opinions supporting this view. Within a likelihood framework, Claeskens and Hjort (2003) proposed a focused information criterion (FIC) that selects the model that minimizes the mean squared error (MSE) of the focus parameter estimator, and demonstrated this FIC's applicability in some common situations. Extensions and generalizations of the FIC have recently appeared in a variety of settings. Hjort and Claeskens (2006) considered the Cox proportional hazard model and developed FICs with respect to four focus parameters: relative risk, cumulative hazard, survival probability, and median survival time. Claeskens, Croux, and Venkerckhoven (2006) introduced new versions of FICs based on L_p loss and error rate in the setting of a logistic regression

Journal of Business & Economic Statistics
January 2012, Vol. 30, No. 1
DOI: 10.1198/jbes.2011.10075

model, whereas Claeskens, Croux, and Venkerckhoven (2007) developed an FIC for the model-order selection problem in an autoregressive model. More recently, Bartolucci and Lupparelli (2008) proposed an FIC for selecting capture-recapture models, and Zhang and Liang (2011) developed an FIC for the generalized additive partial linear model. An excellent discussion of the progress in this area, along with further references, has been provided by Claeskens and Hjort (2008a).

Although the FIC may select a model with better specific properties than the AIC or BIC, whatever model selection criterion is adopted, the true variability involved is always larger than what would have been had the model been given a priori. In reality, the properties of estimators and tests subsequent to model selection depend on how the model is selected, as well as on the model's stochastic nature. Practitioners usually take only the latter into account, however, and report estimates obtained from the chosen model as if they are unconditional even though they are not; consequently, the end results underreport the true variability. [For studies on the properties of estimators subsequent to a range of model selection strategies, see Danilov and Magnus (2004), Leeb and Pötscher (2008), and the references therein.]

An alternative to model selection is model averaging. Rather than selecting a single "winning" model, a model average estimator compromises across a set of competing models. The model average estimator has a distribution that is unconditional on the model selected, and provided that one works with this distribution, inference after model averaging will not suffer from the distortions usually associated with model selection. Another important motivation behind model averaging is that it provides a kind of insurance against selecting a very poor model, and thus holds promise for reducing the risk in regression estimation (Leung and Barron 2006).

There is a large collection of Bayesian literature on model averaging (see, e.g., Hoeting et al. 1999 for a review), but relatively few studies from a frequentist perspective; however, some significant progress has been made in recent years. For instance, Buckland, Burnham, and Augustin (1997) suggested combining models with weights based on the AIC or BIC scores of the competing models. Yang (2001) proposed a frequentist-based adaptive regression mixing method, examined the rate of convergence, and provided simulation evidence on the performance of the resultant estimator. Yuan and Yang (2005) built on this method by implementing a model screening step before combining the models, whereas Hjort and Claeskens (2003) suggested a local misspecification framework for studying the limiting distributions and asymptotic risk properties of post-model selection and model average estimators. More recently, Hansen (2007, 2008) and Wan, Zhang, and Zou (2010) investigated the properties of a frequentist model average estimator based on the Mallows criterion in linear regression. In addition, Magnus, Powell, and Prüfer (2010) and Magnus, Wan, and Zhang (2011) proposed a weighted average least squares estimator, which is a Bayesian mixture of frequentist estimators based on a Laplace prior, Schomaker, Wan, and Heumann (2010) examined the properties of various information criteria-based model average estimators when data are missing, and Zhang and Liang (2011) developed an FIC-based frequentist model average estimator for the generalized additive partial linear model.

In the context of a type I Tobit model with an unknown threshold, in this atrticle we propose FICs for model selection under MSE, mean absolute deviation (MAD), and expected linear exponential (LINEX) loss (Varian 1975; Zellner 1986). As is well known, the choice of loss function often can affect the properties and rankings of estimators. The consideration of different loss criteria thus allows us to examine the extent to which results are robust to departures from a given loss choice. The MAD and MSE are expectations of two common special cases of the L_p loss, and were used by Claeskens and Hjort (2003) and Claeskens, Croux, and Venkerckhoven (2006) for constructing FICs in other contexts. The LINEX loss, which allows for error asymmetries, has been widely used in econometrics research (e.g., Wan and Zou 2003; Elliott and Timmermann 2004; Wan, Zou, and Ohtani 2006; Patton and Timmermann 2007). Recently, Claeskens and Hjort (2008b) applied the LINEX loss in their development of a weighted FIC for model selection in generalized linear models. In addition, Brownlees and Gallo (2011) developed FICs based on expectations of a range of loss functions, including the LINEX loss, for choosing a smoothing parameter in a semiparametric model.

A second objective of this article is to develop model average estimators with values of a smoothed version of the FIC taken as weights. A well-documented potential advantage of model averaging over model selection is a reduction in variance; however, when one is faced with a realistically complex situation with a large number of candidate models, there are inevitable computational difficulties associated with model combining. The development of model screening procedures aimed at removing the poorest-performing models before combining, thereby saving computing time, has been an important development in the recent model averaging literature (e.g., Yuan and Yang 2005; Claeskens, Croux, and Venkerckhoven 2006). In a Monte Carlo study described later, we demonstrate how to implement model screening in conjunction with model averaging in a Tobit model.

The rest of the article is organized as follows. Section 2 presents the model setup, notations, and some basic findings when applying the local misspecification framework to the Tobit model. Section 3 develops FICs under the MSE, MAD, and expected LINEX loss as risk measures. Section 4 provides the results of a Monte Carlo study illustrating the advantages of the FIC over traditional information criteria and demonstrating the merit of model average estimators over model selection estimators. Section 5 applies the proposed criteria and estimators to a dataset on women's working hours, and Section 6 provides some concluding remarks. Proofs of technical results are contained in the Appendix.

2. MODEL FRAMEWORK, NOTATIONS, AND SOME BASIC RESULTS

Consider the following type I Tobit model with a nonzero threshold:

$$y_i^* = x_i'\beta + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{2.1}$$

$$y_i = \begin{cases} y_i^* & \text{if } y_i^* > T, \\ 0 & \text{if } y_i^* \le T, \end{cases} \tag{2.2}$$

where y_i^* is a latent variable, $x_i' = (1, x_{i2}, x_{i3}, \dots, x_{ip})$ is a $1 \times p$ vector of observed regressors, β is a $p \times 1$ vector of parameters, and ε_i is assumed to be iid $\mathrm{N}(0, \sigma^2)$. It is assumed that y_i^* is observed only if $y_i^* > T$, where T is an unknown nonstochastic censoring threshold.

Let $\beta = (\beta_1', \gamma')'$, where β_1 is a $(p-q) \times 1$ vector containing the parameters corresponding to the variables that we want in the model on theoretical or other grounds, and γ is a $q \times 1$ vector of parameters corresponding to the variables that are less certain. In theory, this setup leads to no loss of generality, because β_1 can contain the intercept only if all explanatory variables are uncertain. Following the local misspecification framework of Hjort and Claskens (2003), we define the parameter vector as $\eta = (\theta', \gamma')' = (\theta_0', \gamma_0' + \delta'/n^{1/2})'$, where $\theta = (\sigma, \beta_1')'$ is a $d \times 1$ vector; $d = p - q + 1$; θ_0 is the true value of θ; γ_0, which is usually a null vector, corresponds to the smallest model; and δ is a $q \times 1$ vector of parameters signaling the various degrees of departure from the smallest model. Together, 2^q submodels are obtained by setting different elements of γ to 0. We study the behavior of the estimators when γ deviates from the null vector. For ease of exposition, we do not include the threshold parameter T in the parameter vector η. Following Carson and Sun (2007), T may be estimated by a superconsistent estimator that converges to the true value of T at the rate $1/n$.

Let $\mu = \mu(\eta) = \mu(\theta, \gamma)$ be a parameter estimand, and let $\hat{\theta}_S$ and $\hat{\gamma}_S$ be the maximum likelihood (ML) estimators of θ and γ in the Sth model. Write $\hat{\mu}_S = \mu(\hat{\theta}_S, \hat{\gamma}_S)$. For the purpose of analysis and without loss of generality, we sort the observations such that the first n_0 observations correspond to the censored y_i's and the remaining $n - n_0$ observations correspond to the uncensored y_i's. The likelihood function for the model then may be written as

$$L(\eta) = \prod_{i=1}^{n_0} \Phi\left(\frac{T - x_i'\beta}{\sigma}\right) \prod_{i=n_0+1}^{n} \frac{1}{\sigma} \phi\left(\frac{y_i - x_i'\beta}{\sigma}\right), \tag{2.3}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the probability density function and cumulative distribution function of the $\mathrm{N}(0,1)$ distribution. Now let Y_0 be the $n_0 \times 1$ vector containing the censored y_i's and let Y_1 be the $(n - n_0) \times 1$ vector containing the uncensored y_i's. Also, let X_0 and X_1 be the $n_0 \times p$ and $(n - n_0) \times p$ matrices of observations of explanatory variables corresponding to Y_0 and Y_1. Denote the ith row of X_0 as $x_{i,0}'$, and let $G_{i,i} = T - x_{i,0}'\beta$ and $H_{i,i} = \frac{\phi(G_{i,i}/\sigma)}{\Phi(G_{i,i}/\sigma)}$ be the (i,i)th element of the $n_0 \times n_0$ diagonal matrices G and H. Furthermore, let g and h be $n_0 \times 1$ vectors such that the ith elements of g and h are $g_i = G_{i,i}$ and $h_i = H_{i,i}$. We then obtain the score function

$$\frac{\partial \log L(\eta)}{\partial \eta} = \left(\frac{A - \sigma h'g - \sigma^2 n_1}{\sigma^3}, \left(\frac{B - \sigma X_0'h}{\sigma^2}\right)'\right)' \tag{2.4}$$

and the information matrix

$$J_{n,\mathrm{full}} = -\frac{1}{n}\left(\frac{\partial \log^2 L(\eta)}{\partial \eta\, \partial \eta'}\right) = \begin{pmatrix} J_{n,00} & J_{n,01} \\ J_{n,10} & J_{n,11} \end{pmatrix}_{(d+q)\times(d+q)}$$

$$= -\frac{1}{n}\begin{pmatrix} \frac{g'(2\sigma^2 I_{n_0} - G^2 - \sigma G^2 H)h - 3\sigma A + n_1\sigma^3}{\sigma^5} & \left(\frac{X_0'Ch - 2\sigma B}{\sigma^4}\right)' \\ \frac{X_0'Ch - 2\sigma B}{\sigma^4} & -\frac{X_0'(GH + \sigma H^2)X_0 + \sigma X_1'X_1}{\sigma^3} \end{pmatrix}, \tag{2.5}$$

where I_{n_0} is an $n_0 \times n_0$ identity matrix, $A = (Y_1 - X_1\beta)'(Y_1 - X_1\beta)$, $B = X_1'Y_1 - X_1'X_1\beta$, and $C = \sigma^2 I_{n_0} - G^2 - \sigma GH$. $J_{n,\mathrm{full}}$ is assumed to be of full rank. The elements in the information matrix (2.5) are the key quantities needed for developing the FICs in the next section. Furthermore, we let

$$J_{\mathrm{full}} = \begin{pmatrix} J_{00} & J_{01} \\ J_{10} & J_{11} \end{pmatrix}_{(d+q)\times(d+q)}$$

be the limiting information matrix, with J_{ij} being the limiting value of $J_{n,ij}$ as n approaches infinity, $i, j = 0, 1$.

3. MODEL SELECTION AND MODEL AVERAGING BY FIC

Assume that $\mu = \mu(\theta, \gamma)$ has continuous partial derivatives in a neighborhood of (θ_0, γ_0). Now, from lemma 3.3 of Hjort and Claeskens (2003), we have

$$\sqrt{n}(\hat{\mu}_S - \mu) \xrightarrow{d} \Lambda_S = \left(\frac{\partial \mu}{\partial \theta}\right)' J_{00}^{-1} M + \omega'(\delta - \Psi_S K^{-1} D), \tag{3.1}$$

where $\xrightarrow{d}$ denotes convergence in distribution, $D \sim \mathrm{N}_q(\delta, K)$, $M \sim \mathrm{N}_d(0, J_{00})$ is independent of D, $K = (J_{11} - J_{10}J_{00}^{-1}J_{01})^{-1}$, $\Psi_S = \pi_S'(\pi_S K^{-1}\pi_S')^{-1}\pi_S$, π_S is the projection matrix that maps δ onto the subvector $\delta_S = \pi_S\delta$ that contains the elements of δ in the Sth submodel, and $\omega = J_{10}J_{00}^{-1}\frac{\partial\mu}{\partial\theta} - \frac{\partial\mu}{\partial\gamma}$. It is easy to show that Λ_S follows a normal distribution with mean

$$m_S = E(\Lambda_S) = \omega'(I_q - \Psi_S K^{-1})\delta, \tag{3.2}$$

and variance

$$v_S^2 = \mathrm{Var}(\Lambda_S) = \tau + \omega'\Psi_S K^{-1}\Psi_S\omega, \tag{3.3}$$

where $\tau = (\frac{\partial\mu}{\partial\theta})' J_{00}^{-1} (\frac{\partial\mu}{\partial\theta})$.

It is well known that if T, the censoring threshold, is unknown but coded as 0, then the standard ML estimator of η is inconsistent. Carson and Sun (2007) suggested estimating T by $\hat{T} \equiv \min\{Y_1\}$, which is a superconsistent estimator and converges to the true value of T at the rate $1/n$. They further demonstrated that the ML estimator of η based on $\hat{T}$ is just as efficient as the ML estimator when T is known. In the remainder of the article, we assume that T is estimated by $\hat{T}$. It is straightforward to show that the convergence result given in (3.1) remains valid when T is replaced by $\hat{T}$.

The foregoing results enable derivation of the following FICs:

FIC based on MSE. The aim here is to develop a model selection criterion that chooses the model with the smallest MSE. Now, from (3.1), (3.2), and (3.3), it appears reasonable to use

$$\mathrm{E}(\Lambda_S^2) = \omega'(I_q - \Psi_S K^{-1})\delta\delta'(I_q - \Psi_S K^{-1})'\omega + \omega'\Psi_S K^{-1}\Psi_S\omega + \tau$$

as an approximation to $\mathrm{MSE}(\sqrt{n}\hat{\mu}_S)$. Let $\hat{\delta}$, $\hat{\omega}$, and $\hat{\tau}$ be the ML estimators of δ, ω, and τ using the full model. From Hjort and Claeskens (2003), $\hat{\delta}\hat{\delta}' - K$ is an asymptotically unbiased estimator of $\delta\delta'$. Then, by replacing $\delta\delta'$, K, Ψ_S, ω, and τ by $\hat{\delta}\hat{\delta}' - \hat{K}$, $\hat{K} = (J_{n,11} - J_{n,10}J_{n,00}^{-1}J_{n,01})^{-1}$, $\hat{\Psi}_S = \pi_S'(\pi_S\hat{K}^{-1}\pi_S')^{-1}\pi_S$, $\hat{\omega}$, and $\hat{\tau}$, and removing terms extraneous to the Sth submodel, we obtain the following expression, which we define as the MSE-based FIC for the Sth submodel:

$$\mathrm{FIC}_{\mathrm{MSE}}(S) = (\hat{\omega}'(I_q - \hat{\Psi}_S\hat{K}^{-1})\hat{\delta})^2 + 2\hat{\omega}'\hat{\Psi}_S\hat{\omega}. \tag{3.4}$$

Now, following Hjort and Claeskens (2003), it can be shown that the AIC for the Sth submodel may be written as

$$\mathrm{AIC}_S = -\hat{\delta}'\hat{K}^{-1}\hat{\Psi}_S\hat{K}^{-1}\hat{\delta} + 2|S|,$$

where $|S|$ is the number of uncertain parameters in the Sth submodel. Claeskens and Hjort (2003) showed that when the estimand is $\mu = \log f(y, \theta, \gamma)$ such that $f(y, \theta, \gamma)$ is the probability density function of the data, the MSE-based FIC is asymptotically equivalent to the AIC.

FIC based on MAD. We show in the online supplemental material that

$$\mathrm{E}|\Lambda_S| = 2m_S\left(\Phi\left(\frac{m_S}{v_S}\right) - \frac{1}{2}\right) + 2v_S\phi\left(\frac{m_S}{v_S}\right), \tag{3.5}$$

may be used to approximate $\mathrm{MAD}(\sqrt{n}\hat{\mu}_S)$. Now, by replacing the unknowns in (3.5) by their estimates and removing terms extraneous to the Sth model, we obtain the following expression, which we define as the MAD-based FIC measure for the Sth submodel:

$$\mathrm{FIC}_{\mathrm{MAD}}(S) = 2\hat{m}_S\Phi\left(\frac{\hat{m}_S}{\hat{v}_S}\right) + 2\hat{v}_S\phi\left(\frac{\hat{m}_S}{\hat{v}_S}\right) - \hat{m}_S, \tag{3.6}$$

where $\hat{m}_S = \hat{\omega}'(I_q - \hat{\Psi}_S\hat{K}^{-1})\hat{\delta}$ and $\hat{v}_S = \sqrt{\hat{\tau} + \hat{\omega}'\hat{\Psi}_S\hat{K}^{-1}\hat{\Psi}_S\hat{\omega}}$.

FIC based on expected LINEX errors. For any error e_i, the LINEX loss takes the form

$$\mathrm{LINEX}(e_i, a) = \exp(ae_i) - ae_i - 1,$$

where a ($\neq 0$) is a shape parameter that controls the extent of loss asymmetry. A positive (negative) a indicates that positive (negative) errors are more serious than negative (positive) errors of the same magnitude. When a is close to 0, the LINEX loss reduces to the squared error loss. Now let $a_n = a/\sqrt{n}$. We show in the online supplemental material that

$$\mathrm{LINEX}(\Lambda_S, a) = \exp(a_n^2 v_S^2/2 + a_n m_S) - a_n m_S - 1. \tag{3.7}$$

Again, by substituting estimates for their unknowns and removing terms extraneous to the Sth model, we obtain for the Sth model and a given shape parameter a, the following FIC measure:

$$\mathrm{FIC}_{\mathrm{LINEX}}(S) = \exp(a_n^2\hat{v}_S^2/2 + a_n\hat{m}_S) - a_n\hat{m}_S. \tag{3.8}$$

Write $S_{\mathrm{MSE}} = \arg\min_{S\in\{1,\ldots,2^q\}} \mathrm{FIC}_{\mathrm{MSE}}(S)$, $S_{\mathrm{MAD}} = \arg\min_{S\in\{1,\ldots,2^q\}} \mathrm{FIC}_{\mathrm{MAD}}(S)$, and $S_{\mathrm{LINEX}} = \arg\min_{S\in\{1,\ldots,2^q\}} \mathrm{FIC}_{\mathrm{LINEX}}(S)$. Then the estimators of μ based on the submodels S_{MSE}, S_{MAD}, and S_{LINEX} are the FIC-based model selection estimators.

In contrast, a model average estimator smooths across individual candidate models. The model average estimator that we consider takes the form $\tilde{\mu} = \sum_{\mathrm{all}\ S} w_S\hat{\mu}_S$, where w_S is the weight function for the Sth model. The weight functions are required to sum to 1 or else the estimator is inconsistent (Hjort and Claeskens 2003). As for the choice of w_S, Buckland, Burnham, and Augustin (1997) suggested using weights proportional to $\exp(-\mathrm{AIC}(S)/2)$ or $\exp(-\mathrm{BIC}(S)/2)$, that is,

$$w_S = \frac{\exp(-\,\mathrm{xIC}(S)/2)}{\sum_{\mathrm{all}\ S}\exp(-\,\mathrm{xIC}(S)/2)}, \tag{3.9}$$

where $\mathrm{xIC}(S)$ is the AIC or BIC value based on the Sth submodel. Weights in the form of (3.9) based on AIC and BIC are commonly referred to as smoothed-AIC (S-AIC) and smoothed-BIC (S-BIC) weights. Buckland, Burnham, and Augustin (1997) justified (3.9) by pointing out that the weight ratio $\frac{w_{S_1}}{w_{S_2}}$ is the relative penalized likelihood factor in the case of S-AIC and Schwarz's (1978) approximation to the Bayes factor in the case of S-BIC. However, it is difficult to justify (3.9) when the averaging values are based on FIC scores. As an alternative to (3.9), Hjort and Claeskens (2003) suggested the following weight choice mechanism using the MSE-based FIC scores:

$$w_S = \frac{\exp(-\,\mathrm{FIC}_{\mathrm{MSE}}(S)/(\kappa 2\hat{\omega}'\hat{K}\hat{\omega}))}{\sum_{\mathrm{all}\ S}\exp(-\,\mathrm{FIC}_{\mathrm{MSE}}(S)/(\kappa 2\hat{\omega}'\hat{K}\hat{\omega}))}, \tag{3.10}$$

where $\kappa \geq 0$ is an algorithmic parameter bridging from uniform weighting (κ near 0) to "hard" FIC (large κ). Hjort and Claeskens (2003) demonstrated that (3.10) has an empirical Bayes justification. In addition, when $q = 1$ and $\kappa = 1$, the weights in (3.10) reduce to S-AIC weights. Zhang and Liang (2011) also used (3.10) as the weight choice for their model average estimator in the generalized additive partially linear model and set $\kappa = 1$ in their numerical analysis. The same κ value was used in the simulation studies of Hjort and Claeskens (2003, 2006). Following those studies, we also set $\kappa = 1$. Noting that $2\hat{\omega}'\hat{K}\hat{\omega}$ is the same as $\mathrm{FIC}_{\mathrm{MSE}}$ (F), where the symbol F indexes the full model, we propose the following smoothed-FIC (S-FIC) weight:

$$w_{t,S} = \frac{\exp(-\,\mathrm{FIC}_t(S)/\,\mathrm{FIC}_t\,(\mathrm{F}))}{\sum_{\mathrm{all}\ S}\exp(-\,\mathrm{FIC}_t(S)/\,\mathrm{FIC}_t\,(\mathrm{F}))}, \tag{3.11}$$

where $t = \mathrm{MSE}$, MAD, LINEX. When $t = \mathrm{MSE}$, (3.11) is identical to the weight form (3.10) that sets $\kappa = 1$.

4. A MONTE CARLO STUDY

Along with the procedures' asymptotic properties, investigating their properties when used with smaller samples is also desirable. We do this with a Monte Carlo study that compares (a) estimators from models selected by the proposed FICs and those from models selected by two better known criteria, the AIC and BIC; (b) model average estimators that use values of S-FIC as weights and those based on S-AIC and S-BIC weights; and (c) model average estimators with and without model screening, which is intended to remove the poorly performing models before combining. In keeping with the theoretical results developed in the previous section, the risk associated with the various estimators is evaluated in terms of MAD, MSE, and expected LINEX loss.

Our Monte Carlo study is based on (2.1) and (2.2). We let $x_i' = (1, x_{i2}, x_{i3}, \ldots, x_{ip})$, and consider the following specifications for β:

Case 1: $\beta = (1, 1, 1, 1, 1, 1, 1, 1, 1)'$;

Case 2: $\beta = (1, 1, 1, 1, 1, 0.2, 0.2, 0.2, 0.2)'$;
Case 3: $\beta = (1, 1, 1, 1, 1, 1, 1, 0, 0)'$;
Case 4: $\beta = (1, 1, 1, 0, 0, 0, 0, 0, 0)'$;
Case 5: $\beta = (0.9, 1.5, 1.6, 1.7, 1.5, 0.4, 0.3, 0.2, 0.1, 0, 0)'$;
Case 6: $\beta = (1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0)'$;
Case 7: $\beta = (1, 1.8, 1.9, 2.0, 1.2, 1.5, 0.9, 0.8, 0.4, 0.3, 0.1)'$;
Case 8: $\beta = (1, 0.8, 0.9, 0, 0, 0, 0, 0, 0, 0, 0)'$.

The number of coefficients, p, is 9 for cases 1–4 and 11 for cases 5–8. The latter four cases are precisely the specifications used by Yuan and Yang (2005) in their Monte Carlo study; thus for cases 5–8, we follow Yuan and Yang (2005) by generating $x_{i2}, \dots, x_{i11}$ from a Uniform distribution between $[-1, 1]$, whereas for cases 1–4, we set $(x_{i2}, x_{i3}, \dots, x_{i9})' \sim \mathrm{N}(0, \tilde{D})$, where $\tilde{D}$ is a covariance matrix to be specified later. These cases are chosen to represent different scenarios. For cases 1, 2, and 7, the true model is the full model, and consequently, all candidate models except the full model in the model average are underfitted; in contrast, cases 4 and 8 contain many 0s in the coefficient vector, leading to a large number of overfitted models, whereas cases 3, 5, and 6 are the intermediate cases. For cases 1, 3, 4, 6, and 8, all coefficients in the true models are large, whereas for cases 2, 5, and 7, the true models have some small coefficients. For the correlations among the regressors, we consider the following two scenarios:

Scenario A. All regressors are uncorrelated. Specifically, for cases 1–4, $\tilde{D} = I_8$; and for cases 5–8, $x_{ij} \sim \mathrm{Uniform}[-1, 1], j = 2, \dots, 11$, and $x_{i2}, \dots, x_{i11}$ are independent.

Scenario B. The regressors are moderately correlated. Specifically, for cases 1–4, $\tilde{D} = (\tilde{d}_{ab})$, with $\tilde{d}_{ab} = 0.5^{|a-b|}$, $a, b = 1, \dots, p$; and for cases 5–8, $(x_{i2}, \dots, x_{i11})' = P(x^*_{i2}, \dots, x^*_{i11})'$, where $PP' = \tilde{D}$ is defined as in cases 1–4, $x_{ij} \sim \mathrm{Uniform}[-1, 1], j = 2, \dots, 11$, and $x^*_{i2}, \dots, x^*_{i11}$ are independent.

It is assumed that all models contain the intercept, whereas other regressors are uncertain. The threshold parameter T is set to 1.2. Observation samples of $n = 100$ are generated as training data, and an additional $n_{\text{test}} = 500$ observations are used as test (evaluation) data. We calculate σ from $r^2 = \mathrm{var}(x_i'\beta)/(\mathrm{var}(x_i'\beta) + \sigma^2)$, which takes on values of 0.1–0.95. Each part of the experiment is based on $R = 1000$ replications. Tobit estimates of the unknown parameters are obtained by the ML method. The unknown censored threshold T is estimated by its superconsistent estimator $\hat{T}$.

We also examined a model screening step before model combination. Here we adopted the backward elimination procedure presented by Claeskens, Croux, and Venkerckhoven (2006), which begins with the model containing all regressors and removes one variable at a time by deleting the variable that when removed from the current model leads to the smallest value for the information criterion. The procedure continues until $q + 1$ nested models are obtained. Model average estimators of the unknowns are obtained by compromising across these $q + 1$ submodels. From a computational standpoint, this model screening procedure has the practical advantage of reducing the number of models to be combined from 2^q to $q + 1$.

Evaluation of estimator performance is based on the test data. For the jth observation the quantity of interest is $\mu(j) = E(y_j|x_j)$, the predicted value of the response. Straightforward calculations yield

$$\mu(j) = \Phi\left(\frac{x_j'\beta - T}{\sigma}\right)x_j'\beta + \sigma\phi\left(\frac{x_j'\beta - T}{\sigma}\right). \tag{4.1}$$

For the Sth model, the selection estimator $\hat{\mu}_S(j)$ is obtained by replacing β, T, and σ in (4.1) by their estimators obtained from the Sth model. The model average estimator is a weighted average of $\hat{\mu}(j)$'s across either all or a subset of the 2^q candidate models, depending on whether a model screening procedure is implemented.

The MAD, MSE, and expected LINEX loss of $\hat{\mu}_S$ over all observations in the test data are computed as

$$\mathrm{MAD}(\hat{\mu}_S) = \frac{1}{R}\sum_{r=1}^{R}\frac{1}{n_{\text{test}}}\sum_{j=1}^{n_{\text{test}}}|\hat{\mu}_S^r(j) - \mu(j)|,$$

$$\mathrm{MSE}(\hat{\mu}_S) = \frac{1}{R}\sum_{r=1}^{R}\frac{1}{n_{\text{test}}}\sum_{j=1}^{n_{\text{test}}}(\hat{\mu}_S^r(j) - \mu(j))^2,$$

and

$$\mathrm{LINEX}(\hat{\mu}_S) = \frac{1}{R}\sum_{r=1}^{R}\frac{1}{n_{\text{test}}}\sum_{j=1}^{n_{\text{test}}}\big(\exp\big[a(\hat{\mu}_S^r(j) - \mu(j))\big] - a(\hat{\mu}_S^r(j) - \mu(j)) - 1\big),$$

where $\hat{\mu}_S^r$ is the estimator of μ obtained from the Sth model in the rth replication of the experiment. Note that the Sth model simply indexes the model chosen by a particular selection criterion. With a given criterion, the actual chosen model varies across the replications of the experiment; that is, the Sth model is not fixed in repeated samples. For the expected LINEX loss, we choose $a = -1.5$, 0.5, and 1.5, representing different degrees of negative as well as positive loss asymmetry. The MAD, MSE, and expected LINEX loss of the model average estimators $\tilde{\mu}$ are defined analogously. To assess the precision of these risk measures, we also calculate their standard errors (se). For example, the se of the MAD of $\hat{\mu}_S$ is calculated by

$$\mathrm{se}(\mathrm{MAD}(\hat{\mu}_S)) = \left(\frac{1}{R-1}\sum_{r=1}^{R}\left(\mathrm{MAD}_r - \frac{1}{R}\sum_{r=1}^{R}\mathrm{MAD}_r\right)^2\right)^{1/2} \Big/ \sqrt{R},$$

where $\mathrm{MAD}_r = \frac{1}{n_{\text{test}}}\sum_{j=1}^{n_{\text{test}}}|\hat{\mu}_S^r(j) - \mu(j)|$.

To save space, we only provide the results for Scenario A of case 1 and Scenario B of case 8 in Tables 1 and 2. Results of other cases are given in Tables R1–R14 in Section B of the online supplemental material. In each table, panel A gives the results based on a submodel selected by a model selection criterion, whereas panels B and C give the model averaged results without and with model screening. The results reported under the FIC category pertain to the FIC constructed under the stated performance criterion. For example, in panel A, when reporting the MAD for the FIC, it is assumed that the submodel selected is the one with the smallest $\mathrm{FIC}_{\mathrm{MAD}}$; the results corresponding to S-FIC in panels B and C are interpreted similarly. Given that

Table 1. Simulation results for case 1 (Scenario A)

r^2	Criterion	MAD (se)	MSE (se)	LINEX (se) $a=-1.5$	LINEX (se) $a=0.5$	LINEX (se) $a=1.5$
	Panel A: Model selection					
0.2	AIC	1.128 (0.019)	2.302 (0.083)	24.438 (4.455)	0.436 (0.036)	925.264 (445.179)*
	BIC	1.338 (0.012)*	3.132 (0.059)*	69.670 (8.649)*	0.538 (0.034)*	608.827 (199.110)
	FIC	0.936 (0.024)†	1.598 (0.076)†	5.773 (1.010)†	0.230 (0.016)†	4.412 (0.710)†
	Panel B: Model averaging without screening					
	S-AIC	1.009 (0.019)	1.857 (0.073)	13.246 (2.168)	0.314 (0.026)	319.826 (183.958)*
	S-BIC	1.101 (0.016)*	2.157 (0.067)*	27.025 (3.662)*	0.324 (0.023)*	235.405 (133.375)
	S-FIC	**0.823** (0.019)†	**1.303** (0.057)†	5.459 (0.836)†	**0.137** (0.005)†	**1.078** (0.064)†
	Panel C: Model averaging with screening					
	S-AIC	1.062 (0.018)	2.041 (0.074)	15.004 (2.735)	0.368 (0.030)	489.801 (239.168)*
	S-BIC	1.204 (0.014)*	2.555 (0.063)*	40.694 (5.175)*	0.408 (0.028)*	477.003 (261.795)
	S-FIC	0.917 (0.022)†	1.574 (0.072)†	**4.364** (0.651)†	0.183 (0.011)†	1.463 (0.126)†
	Panel A: Model selection					
0.4	AIC	0.732 (0.022)	1.096 (0.065)	4.872 (1.016)	0.164 (0.017)	60.824 (55.568)*
	BIC	1.024 (0.021)*	2.015 (0.075)*	47.887 (7.524)*	0.295 (0.018)*	56.015 (36.470)
	FIC	0.613 (0.015)†	0.766 (0.041)†	1.528 (0.213)†	0.097 (0.008)†	1.433 (0.449)†
	Panel B: Model averaging without screening					
	S-AIC	0.744 (0.017)	1.084 (0.048)	5.044 (0.734)	0.150 (0.011)	16.962 (13.376)*
	S-BIC	0.898 (0.017)*	1.528 (0.055)*	18.632 (2.766)*	0.201 (0.012)*	15.851 (10.215)
	S-FIC	0.622 (0.015)†	**0.761** (0.039)†	1.747 (0.275)†	0.112 (0.003)†	0.764 (0.024)†
	Panel C: Model averaging with screening					
	S-AIC	0.742 (0.017)	1.083 (0.051)	4.445 (0.661)	0.157 (0.013)	28.214 (23.946)
	S-BIC	0.956 (0.018)*	1.735 (0.062)*	27.486 (4.392)*	0.242 (0.015)*	37.986 (28.458)*
	S-FIC	**0.607** (0.016)†	0.770 (0.041)†	**1.346** (0.184)†	**0.088** (0.005)†	**0.728** (0.052)†
	Panel A: Model selection					
0.6	AIC	**0.436** (0.013)†	**0.416** (0.026)†	0.723 (0.088)†	0.056 (0.004)†	0.956 (0.141)
	BIC	0.586 (0.024)*	0.795 (0.060)*	3.852 (1.579)*	0.106 (0.008)*	2.872 (0.520)*
	FIC	0.458 (0.011)	0.462 (0.020)	0.873 (0.092)	0.056 (0.003)†	0.547 (0.041)†
	Panel B: Model averaging without screening					
	S-AIC	0.483 (0.013)	0.502 (0.027)	1.103 (0.132)	0.066 (0.004)†	1.094 (0.112)
	S-BIC	0.591 (0.018)*	0.751 (0.042)*	3.093 (0.514)*	0.097 (0.006)	1.954 (0.253)*
	S-FIC	0.475 (0.010)†	0.433 (0.019)†	0.727 (0.063)†	0.102 (0.002)*	0.649 (0.011)†
	Panel C: Model averaging with screening					
	S-AIC	0.470 (0.013)	0.477 (0.025)	0.961 (0.108)	0.064 (0.004)	1.063 (0.110)
	S-BIC	0.587 (0.019)*	0.751 (0.045)*	2.741 (0.463)*	0.099 (0.006)*	2.223 (0.339)*
	S-FIC	0.438 (0.010)†	0.429 (0.019)†	**0.627** (0.063)†	**0.052** (0.002)†	**0.426** (0.017)†
	Panel A: Model selection					
0.8	AIC	**0.280** (0.007)†	**0.182** (0.009)†	**0.258** (0.017)†	**0.023** (0.001)†	**0.254** (0.016)†
	BIC	0.286 (0.009)	0.196 (0.015)	**0.258** (0.017)†	0.025 (0.002)	0.309 (0.047)
	FIC	0.404 (0.009)*	0.380 (0.015)*	0.751 (0.047)*	0.045 (0.002)*	0.401 (0.022)*
	Panel B: Model averaging without screening					
	S-AIC	0.284 (0.007)†	0.188 (0.010)†	0.271 (0.020)†	0.024 (0.001)†	0.266 (0.020)†
	S-BIC	0.291 (0.008)	0.201 (0.013)	0.317 (0.039)	0.026 (0.002)	0.299 (0.032)
	S-FIC	0.376 (0.008)*	0.284 (0.012)*	0.491 (0.018)*	0.109 (0.002)*	0.649 (0.008)*
	Panel C: Model averaging with screening					
	S-AIC	0.283 (0.007)†	0.188 (0.010)†	0.270 (0.020)†	0.024 (0.001)†	0.265 (0.019)†
	S-BIC	0.290 (0.008)	0.200 (0.013)	0.311 (0.037)	0.026 (0.002)	0.297 (0.032)
	S-FIC	0.343 (0.007)*	0.287 (0.012)*	0.388 (0.021)*	0.039 (0.001)*	0.302 (0.009)*

these tables contain voluminous numerical details, to facilitate comparisons, the lowest and largest risk figures in each panel corresponding to a given r^2 value are indicated by "†" and "*," respectively, and the lowest risk figure across all three panels is highlighted in bold type.

The following conclusions are drawn from the full results, given in Tables 1 and 2, as well as in Tables R1–R14 in the supplemental material:

(1) Except for the small model scenario represented by cases 4 and 8, FIC model selection generally results in more

Table 1. (Continued)

r^2	Criterion	MAD (se)	MSE (se)	LINEX (se) $a=-1.5$	LINEX (se) $a=0.5$	LINEX (se) $a=1.5$
	Panel A: Model selection					
0.95	AIC	**0.150** (0.004)†	**0.062** (0.004)†	**0.080** (0.006)†	**0.008** (0.001)†	**0.078** (0.006)†
	BIC	**0.150** (0.004)†	**0.062** (0.004)†	**0.080** (0.006)†	**0.008** (0.001)†	**0.078** (0.006)†
	FIC	0.440 (0.007)*	0.492 (0.014)*	1.217 (0.054)*	0.107 (0.003)*	1.622 (0.110)*
	Panel B: Model averaging without screening					
	S-AIC	**0.150** (0.004)†	**0.062** (0.004)†	**0.080** (0.006)†	**0.008** (0.001)†	**0.078** (0.006)†
	S-BIC	**0.150** (0.004)†	**0.062** (0.004)†	**0.080** (0.006)†	**0.008** (0.001)†	**0.078** (0.006)†
	S-FIC	0.372 (0.006)*	0.350 (0.011)*	0.545 (0.014)*	0.121 (0.001)*	0.855 (0.011)*
	Panel C: Model averaging with screening					
	S-AIC	**0.150** (0.004)†	**0.062** (0.004)†	**0.080** (0.006)†	**0.008** (0.001)†	**0.078** (0.006)†
	S-BIC	**0.150** (0.004)†	**0.062** (0.004)†	**0.080** (0.006)†	**0.008** (0.001)†	**0.078** (0.006)†
	S-FIC	0.322 (0.006)*	0.279 (0.010)*	0.477 (0.023)*	0.048 (0.001)*	0.399 (0.011)*

efficient estimates than the traditional AIC- and BIC- based selection with respect to all dimensions of performance when r^2 is small (say $\leq$0.2); however, the reverse is observed when r^2 is large (say $\geq$0.8). This indicates that FIC-based model selection is likely to be advantageous when the underlying true model is large and has a high variance noise level; however, its usefulness may be questionable in other circumstances. Our results also suggest that for cases where the FIC results in the best of the three model selection strategies, it is usually under expected LINEX loss with $a = 1.5$ or $a = -1.5$ that the improvement in efficiency from using FIC over AIC and BIC is most pronounced. Except for cases 4 and 8, the risks of the FIC model selection estimator under all three evaluation criteria generally cross those of the AIC and BIC model selection estimators at moderate r^2 values.

(2) For cases 1, 2, and 7, where the true models are large and have no zero coefficient, BIC is the least efficient of the three model selection strategies most of the time, whereas AIC is almost always the most efficient when r^2 is large, and, as discussed earlier, FIC generally has an advantage over the other two strategies when r^2 is small. However, as the number of zero coefficients increases (as in cases 3, 5, and 6), we find evidence that BIC outperforms AIC and/or FIC more frequently, especially for $r^2 \geq 0.6$, even though for small r^2, BIC is still often the least favored. For cases 4 and 8, where the underlying true models have a large number of zero coefficients, the BIC estimator habitually has the smallest risk among the three model selection estimators across all r^2 values. This result is expected given that cases 4 and 8 represent the small model scenario and that BIC favors parsimony. Interestingly, when the true model is small, the FIC can result in the least accurate model selection estimates even for small r^2 values. Thus it seems that the FIC is most credible when the underlying true model is large with unstable noises, but is least suitable for small models that are relatively stable in variance.

(3) Although neither model averaging nor model selection is always a better strategy than the other, the former is found to be the preferred strategy in a high proportion of circumstances. When r^2 is small (say $\leq$0.2), the most accurate estimates are invariably obtained by model averaging. This result is not unexpected, because the large error variance associated with the true model when r^2 is small makes it difficult to identify the best model, thus making model averaging, which shields against choosing a very bad model, a more viable strategy than selecting a single model. Yuan and Yang (2005) reached a similar conclusion when comparing their adaptive regression model mixing method with model selection. When r^2 is large (say $\geq$0.8), AIC- and BIC-based model selection estimators can deliver more accurate estimates than their model averaging counterparts, but when that occurs, these model selection and average estimators generally have very similar levels of accuracy. In fact, our Monte Carlo results suggest that when r^2 is large, there often is very little to choose from among the six AIC- and BIC-based model selection and average estimators. In contrast, for the FIC-based estimators, model averaging generally results in a more noticeable improvement in efficiency, but even with this reduction, the FIC model average estimators rarely outperform the AIC- and BIC-based estimators when r^2 is large.

(4) When r^2 is small, model screening rarely improves the performance of a given model average estimator. When $r^2 \geq$ 0.4, the comparison between the screening and nonscreening versions of the model average estimator is not as clear-cut. For these cases, the nonscreening version still appears to have the edge most of the time when model averaging is based on AIC or BIC, but the screening version also can produce superior estimates in many circumstances, especially when FIC is used as the basis for averaging. Perhaps a more important finding is that when model averaging without screening yields the most precise estimates, typically the superiority of model averaging over model selection is not reversed by implementing screening on averaging, although there are some clear exceptions (see, e.g., the MSE comparisons of the FIC-based estimators in Table 1 with $r^2 = 0.4$). This finding is encouraging in view of the fact that model screening significantly reduces the number of candidate models to be combined.

(5) Although there are exceptions, the ordinal rankings of the S-AIC, S-BIC, and S-FIC model average estimators, with and without screening, generally follow the same pattern as the ordinal rankings of their model selection counterparts. In all scenarios except the small model scenario, S-FIC model averaging usually yields better estimates than S-AIC and S-BIC averag-

Table 2. Simulation results for case 8 (Scenario B)

				LINEX (se)		
r^2	Criterion	MAD (se)	MSE (se)	$a = -1.5$	$a = 0.5$	$a = 1.5$
	Panel A: Model selection					
0.2	AIC	0.350 (0.011)	0.215 (0.013)	0.265 (0.019)	0.028 (0.002)	0.316 (0.026)*
	BIC	0.306 (0.009)†	0.164 (0.010)†	0.206 (0.014)†	0.021 (0.001)†	0.218 (0.017)†
	FIC	0.383 (0.009)*	0.231 (0.011)*	0.306 (0.016)*	0.031 (0.002)*	0.310 (0.022)
	Panel B: Model averaging without screening					
	S-AIC	0.305 (0.008)*	0.159 (0.008)*	0.191 (0.011)*	0.021 (0.001)*	0.216 (0.016)*
	S-BIC	**0.276** (0.008)†	**0.130** (0.007)†	**0.163** (0.010)†	**0.016** (0.001)†	0.162 (0.011)
	S-FIC	0.303 (0.007)	0.157 (0.007)	0.185 (0.009)	0.017 (0.001)	**0.153** (0.007)†
	Panel C: Model averaging with screening					
	S-AIC	0.329 (0.010)	0.189 (0.011)	0.227 (0.015)	0.025 (0.002)*	0.274 (0.022)*
	S-BIC	0.290 (0.009)†	0.148 (0.009)†	0.189 (0.014)†	0.019 (0.001)†	0.189 (0.015)†
	S-FIC	0.365 (0.009)*	0.212 (0.010)*	0.254 (0.013)*	0.025 (0.001)*	0.241 (0.015)
	Panel A: Model selection					
0.4	AIC	0.231 (0.009)	0.105 (0.007)	0.125 (0.010)	0.013 (0.001)	0.138 (0.011)
	BIC	**0.203** (0.010)†	0.088 (0.008)†	0.112 (0.012)†	0.011 (0.001)†	0.109 (0.011)†
	FIC	0.284 (0.006)*	0.143 (0.006)*	0.190 (0.010)*	0.018 (0.001)*	0.164 (0.008)*
	Panel B: Model averaging without screening					
	S-AIC	0.221 (0.007)	0.092 (0.006)	0.110 (0.008)	0.012 (0.001)	0.117 (0.009)*
	S-BIC	0.205 (0.008)†	**0.082** (0.006)†	**0.101** (0.008)†	**0.010** (0.001)†	**0.099** (0.007)†
	S-FIC	0.232 (0.005)*	0.102 (0.005)*	0.125 (0.007)*	0.013 (0.001)*	0.107 (0.004)
	Panel C: Model averaging with screening					
	S-AIC	0.228 (0.008)	0.099 (0.006)	0.117 (0.009)	0.013 (0.001)	0.128 (0.010)
	S-BIC	0.205 (0.008)†	0.084 (0.006)†	0.104 (0.009)†	0.011 (0.001)†	0.102 (0.008)†
	S-FIC	0.262 (0.006)*	0.121 (0.006)*	0.141 (0.007)*	0.014 (0.001)*	0.130 (0.007)*
	Panel A: Model selection					
0.6	AIC	0.180 (0.007)	0.066 (0.005)	0.079 (0.006)	0.008 (0.001)	0.080 (0.006)
	BIC	**0.138** (0.006)†	0.040 (0.003)†	**0.046** (0.004)†	**0.005** (0.000)†	0.047 (0.005)†
	FIC	0.240 (0.005)*	0.117 (0.005)*	0.169 (0.010)*	0.014 (0.001)*	0.116 (0.005)*
	Panel B: Model averaging without screening					
	S-AIC	0.164 (0.005)	0.053 (0.004)	0.063 (0.005)	0.007 (0.000)	0.063 (0.005)
	S-BIC	0.140 (0.005)†	**0.039** (0.003)†	0.047 (0.004)†	**0.005** (0.000)†	**0.045** (0.003)†
	S-FIC	0.194 (0.005)*	0.077 (0.004)*	0.108 (0.005)*	0.013 (0.000)*	0.098 (0.003)*
	Panel C: Model averaging with screening					
	S-AIC	0.174 (0.006)	0.060 (0.004)	0.072 (0.005)	0.008 (0.001)	0.072 (0.005)
	S-BIC	0.140 (0.005)†	**0.039** (0.003)†	**0.046** (0.004)†	**0.005** (0.000)†	**0.045** (0.004)†
	S-FIC	0.203 (0.005)*	0.079 (0.004)*	0.093 (0.006)*	0.009 (0.000)*	0.081 (0.004)*
	Panel A: Model selection					
0.8	AIC	0.117 (0.004)	0.032 (0.002)	0.038 (0.003)	0.004 (0.000)	0.036 (0.002)
	BIC	**0.088** (0.004)†	0.019 (0.002)†	0.023 (0.002)†	**0.002** (0.000)†	0.021 (0.002)†
	FIC	0.197 (0.004)*	0.107 (0.004)*	0.173 (0.008)*	0.012 (0.000)*	0.093 (0.003)*
	Panel B: Model averaging without screening					
	S-AIC	0.105 (0.003)	0.025 (0.002)	0.030 (0.002)	0.003 (0.000)	0.028 (0.002)
	S-BIC	**0.088** (0.003)†	**0.017** (0.001)†	**0.021** (0.002)†	**0.002** (0.000)†	**0.019** (0.001)†
	S-FIC	0.154 (0.004)*	0.056 (0.003)*	0.111 (0.004)*	0.015 (0.000)*	0.106 (0.002)*
	Panel C: Model averaging with screening					
	S-AIC	0.111 (0.004)	0.028 (0.002)	0.034 (0.003)	0.004 (0.000)	0.032 (0.002)
	S-BIC	**0.088** (0.003)†	**0.017** (0.001)†	**0.021** (0.002)†	**0.002** (0.000)†	**0.019** (0.002)†
	S-FIC	0.137 (0.004)*	0.042 (0.002)*	0.051 (0.003)*	0.005 (0.000)*	0.045 (0.002)*

ing when r^2 is small, and, conversely, also when r^2 is large. For cases 4 and 8, where the true models are small, S-BIC model averaging is habitually preferred over AIC- or FIC-based model averaging across all r^2 values.

(6) All things being equal, estimators' risks are generally larger under Scenario B, where the regressors are moderately correlated, than under Scenario A, where the regressors are uncorrelated. Typically, correlations among regressors do not af-

Table 2. (Continued)

r^2	Criterion	MAD (se)	MSE (se)	LINEX (se) $a=-1.5$	LINEX (se) $a=0.5$	LINEX (se) $a=1.5$
	Panel A: Model selection					
0.95	AIC	0.068 (0.002)	0.015 (0.001)	0.018 (0.001)	0.002 (0.000)	0.016 (0.001)
	BIC	**0.048** (0.002)†	0.008 (0.001)†	**0.009** (0.001)†	**0.001** (0.000)†	**0.008** (0.001)†
	FIC	0.161 (0.003)*	0.106 (0.003)*	0.229 (0.006)*	0.022 (0.000)*	0.210 (0.005)*
	Panel B: Model averaging without screening					
	S-AIC	0.058 (0.002)	0.011 (0.001)	0.013 (0.001)	**0.001** (0.000)†	0.012 (0.001)
	S-BIC	**0.048** (0.002)†	**0.007** (0.001)†	**0.009** (0.001)†	**0.001** (0.000)†	**0.008** (0.001)†
	S-FIC	0.115 (0.002)*	0.038 (0.001)*	0.134 (0.002)*	0.017 (0.000)*	0.130 (0.001)*
	Panel C: Model averaging with screening					
	S-AIC	0.063 (0.002)	0.013 (0.001)	0.015 (0.001)	0.002 (0.000)	0.014 (0.001)
	S-BIC	**0.048** (0.002)†	**0.007** (0.001)†	**0.009** (0.001)†	**0.001** (0.000)†	**0.008** (0.001)†
	S-FIC	0.080 (0.002)*	0.021 (0.001)*	0.029 (0.001)*	0.004 (0.000)*	0.029 (0.001)*

fect the relative performance of the BIC-based estimators when these estimators are ranked best (e.g., under cases 4 and 8), but in other situations, regressor correlations do appear to have an impact on the relative performance of estimators, and estimators' rankings can change as one goes from Scenario A to Scenario B. In addition, there are fewer instances in which model selection is preferred over model averaging under Scenario B than under Scenario A. One possible explanation for this is that correlation increases uncertainty, making it more difficult to identify the best model accurately.

(7) Commonly, the rankings of estimators are invariant with respect to the performance yardsticks; when the rankings of the estimators are not consistent across the different evaluation criteria, it is usually the ranking based on expected LINEX loss that is at variance with those based on MAD and MSE, although occasionally the rankings produced by the latter two measures can differ as well. Thus, it appears that the relative performance of the various estimators is reasonably robust to the choice of evaluation functions. The difference in rankings occasionally produced by LINEX is expected given that this loss function is asymmetric, with very different properties from the symmetric MAD and MSE, especially when $|a|$ is large, which signals a strong degree of asymmetry.

5. ANALYSIS OF REAL DATA

In this section we consider a practical application of the proposed methods using Mroz's (1987) well-known study on female labor supply, based on a 1975 dataset that consists of $n = 752$ married white women age 30–65, 427 of whom worked at least part time during the year. The censored response variable is a woman's total work hours during the year. The explanatory variables, listed in Table 3, contain information on each woman, including age, educational level, years of work experience, spouse's years of work experience, number of children, and different age groups. A Tobit model with a nonzero threshold in the form of (2.1) and (2.2) is used to analyze this dataset. Table 3 presents the ML estimates of the coefficients and estimate of the threshold parameter based on the super-consistent estimator proposed by Carson and Sun (2007). The pseudo-R^2 value reported is a goodness-of-fit measure for the limited dependent variables proposed by Laitila (1993).

Our evaluation of the aforementioned model selection and averaging methods is based on leave-k-out cross-validation, through which each partitioned set contains 400 random observations as training data and leaves out the remaining $k = 352$ observations as test data. This process is repeated 500 times. The performance of the various methods is then gauged based on the average accuracy of the forecasts that they produce for the censored response variable across the 500 replications. Again, model averaging is conducted both with and without model screening, with the former resulting in an estimator that compromises across 9 models, and the latter being a weighted average of estimators in 256 models.

Table 4 presents the empirical MAD, MSE, and expected LINEX loss of the forecasts and their standard errors produced by the various methods across the 500 replications. These quantities are calculated in much the same manner as the corresponding quantities in the Monte Carlo study. Again, the best and worst risk values in a given panel are indicated by "†" and "*," respectively, and the best estimate across all the three panels is highlighted in bold type. We set a, the shape parameter of the LINEX loss function, to -0.01, -0.0055, 0.0055, and 0.01. Notwithstanding these relatively small $|a|$ values, the degrees of asymmetry are not necessarily mild, because the errors in the present analysis are generally not small. We also have found that values of $|a| > 0.01$ tend to result in very large estimator risk values, making comparisons meaningless.

The results reported in Table 4 are somewhat mixed regarding the relative performance of the various estimators. Under MSE and expected LINEX loss, FIC typically results in the most efficient estimators for both model selection and model averaging, and there are clear advantages of using FIC-based estimators over their AIC and BIC counterparts. An exception occurs when a in the LINEX loss is set to -0.0055, for which the FIC-based model average estimator without screening performs worse than its AIC and BIC counterparts. Under MAD, FIC-based estimators generally perform poorly, and the non-screening version of the FIC model average estimator has the worst performance of all estimators across the three panels. Table 4 also shows that except when for $a = -0.01$ and -0.0055

Table 3. List of explanatory variables and estimation results for the female labor supply example

Explanatory variables	ML estimates of coefficients ($\times 10^3$)	t-ratios
Intercept	2.178	4.739
Age of woman	−0.057	−4.368
Education years of woman	0.089	3.478
Working years of woman	0.076	11.637
Number of kids younger than 6	−0.972	−8.527
Number of kids between 6 and 18	−0.039	−0.989
Age of husband	−0.011	−0.898
Education years of husband	−0.034	−1.700
Income excluding woman's wage ($1000)	−0.006	−1.274
$\hat{\sigma} = 1.126 \times 10^3$, $\hat{T} = 12$, Pseudo-$R^2 = 0.345$, Log-likelihood $= 5.602 \times 10^5$		

under LINEX loss, the best performance across the three panels is achieved by model averaging, with the nonscreening version of the S-AIC estimator being the best in terms of MAD and either the screening or nonscreening version of the S-FIC estimator being the best in terms of MSE and expected LINEX loss with $a = 0.0055$ and 0.01. In contrast; on the other hand, when a is set to -0.01 or -0.0055, the FIC-based model selection estimator is the preferred estimator across the three panels. Thus, if focus is centered on a particular information criterion, then model averaging does not always produce superior estimates. Table 4 also shows that screening can both improve and worsen the performance of model average estimators. In broad terms, these results are consistent with the Monte Carlo findings reported in Section 4. Although no single case of our Monte Carlo experiments produces results that exactly match the results reported here, many of the features seen here are also evident from the Monte Carlo findings. Also, from Table 3, the pseudo-R^2 value for the current regression model is 0.345. If this sample statistic is a reasonable reflection of the true population variability, then it is not surprising that the results neither strongly support nor not support FIC and model averaging. Recall that in the "intermediate zone" where data variability is moderate, the risks of the various FIC and model average estimators tend to cross those of their corresponding AIC, BIC, and model selection counterparts. Even so, the results from this real data example clearly indicate that there are many circumstances in which FIC and model averaging can lead to efficiency gains over the traditional AIC and BIC model selection procedures.

6. CONCLUDING REMARKS

This article has considered model selection and model averaging in the Tobit setting. Although our results offer some interesting insights into the application of the Tobit model, it also raises some important issues that merit further study. For example, there is a need to analyze the distribution of the model average estimator considered here. As we have emphasized, model averaging does not guarantee that subsequent inference will be on sound footing unless based on the distribution of the model average estimator. A fruitful avenue for further research would be the derivation of the model average estimator's full distribution. The recent related work of Pötscher (2006) may serve as a useful guide in this regard.

The analyses and results reported here pertain only to a type I Tobit model. Extensions of investigations into other common types of Tobit models appear to be desirable for both theoretical and practical reasons. Moreover, the model averaging scheme developed in this article is just one among a number of suggested methods of combining estimators in the rapidly expanding literature on frequentist model averaging. Other model

Table 4. Estimator performance for the female labor supply example

Criterion	MAD (se) ($\times 10^2$)	MSE (se) ($\times 10^5$)	LINEX (se) ($\times 10^{12}$) $a = -0.01$	LINEX (se) ($\times 10^5$) $a = -0.0055$	LINEX (se) ($\times 10^2$) $a = 0.0055$	LINEX (se) ($\times 10^8$) $a = 0.01$
Panel A: Model selection						
AIC	5.865 (0.008)†	5.805 (0.020)	2.450 (0.132)*	4.622 (0.161)*	9.185 (2.321)	4.026 (3.739)
BIC	5.882 (0.008)*	5.811 (0.020)*	1.998 (0.109)	4.200 (0.145)	11.461 (3.324)*	7.782 (7.260)*
FIC	5.879 (0.008)	5.792 (0.020)†	**1.648** (0.102)†	**3.906** (0.139)†	5.662 (0.453)†	0.075 (0.017)†
Panel B: Model averaging without screening						
S-AIC	**5.858** (0.008)†	5.782 (0.020)	2.138 (0.113)*	4.359 (0.150)	8.912 (1.308)	1.567 (1.097)
S-BIC	5.873 (0.007)	5.785 (0.020)*	1.934 (0.103)†	4.153 (0.142)†	10.920 (2.215)*	3.928 (3.220)*
S-FIC	5.943 (0.007)*	**5.769** (0.020)†	1.995 (0.119)	4.360 (0.181)*	**1.728** (0.137)†	**0.002** (0.000)†
Panel C: Model averaging with screening						
S-AIC	**5.858** (0.008)†	5.788 (0.020)*	2.267 (0.121)*	4.471 (0.155)*	8.164 (1.184)	1.282 (0.953)
S-BIC	5.870 (0.008)	5.784 (0.020)	1.934 (0.102)	4.156 (0.142)	10.670 (2.977)*	6.372 (5.929)*
S-FIC	5.875 (0.008)*	5.782 (0.020)†	1.823 (0.103)†	4.085 (0.151)†	4.582 (0.408)†	0.006 (0.001)†

mixing schemes, such as that based on the weighted average least squares method developed recently by Magnus, Powell, and Prüfer (2010) and Magnus, Wan, and Zhang (2011), which combines frequentist estimators by Laplace priors, may be adopted and generalized to a Tobit setting and then compared with the model averaging scheme considered here.

SUPPLEMENTARY MATERIALS

Proofs and results: The online supplemental material contains the proofs of Equations (3.5) and (3.7) and the results of the Monte Carlo study. (supplements.pdf)

ACKNOWLEDGMENTS

The authors thank the editor and referees for their very helpful suggestions that greatly improved the article, and Ejaz Ahmed, Peter Cribbett, John Hinde, and Guohua Zou for their comments. Xinyu Zhang's work was supported by the National Natural Science Foundation of China (grants 70625004 and 70933003). Alan Wan's and Sherry Zhou's work was supported by the Hong Kong Research Grants Council (grants CityU-102709 and CityU-150809).

[Received March 2010. Revised June 2011.]

REFERENCES

Amemiya, T. (1981), "Qualitative Response Models: A Survey," *Journal of Economic Literature*, 19, 1483–1536. [132]

——— (1984), "Tobit Models: A Survey," *Journal of Econometrics*, 24, 3–61. [132]

——— (1985), *Advanced Econometrics*, Oxford: Blackwell. [132]

Bartolucci, F., and Lupparelli, M. (2008), "Focused Information Criterion for Capture-Recapture Models for Closed Populations," *Scandinavian Journal of Statistics*, 35, 629–649. [133]

Brownlees, C. T., and Gallo, G. M. (2011), "Shrinkage Estimation of Semiparametric Multiplicative Error Models," *International Journal of Forecasting*, 27, 365–378. [133]

Buckland, S. T., Burnham, K. P., and Augustin, N. H. (1997), "Model Selection: An Integral Part of Inference," *Biometrics*, 53, 603–618. [133,135]

Carson, R. T., and Sun, Y. (2007), "The Tobit Model With Non-Zero Threshold," *Econometrics Journal*, 10, 488–502. [134,140]

Claeskens, G., and Hjort, N. L. (2003), "The Focused Information Criterion" (with discussion), *Journal of the American Statistical Association*, 98, 900–945. [132,133,135]

——— (2008a), *Model Selection and Model Averaging*, New York: Cambridge University Press. [132,133]

——— (2008b), "Minimizing Average Risk in Regression Models," *Econometric Theory*, 24, 493–527. [132,133]

Claeskens, G., Croux, C., and Venkerckhoven, J. (2006), "Variable Selection for Logit Regression Using a Prediction-Focused Information Criterion," *Biometrics*, 62, 972–979. [132,133,136]

——— (2007), "Prediction-Focused Model Selection for Autoregressive Models," *Australian and New Zealand Journal of Statistics*, 49, 972–979. [133]

Danilov, D., and Magnus, J. R. (2004), "On the Harm That Ignoring Pretesting Can Cause," *Journal of Econometrics*, 122, 27–46. [133]

Dhrymes, P. J. (1986), "Limited Dependent Variables," in *Handbook of Econometrics 3*, eds. Z. Grilliches and M. D. Intriligator, Amsterdam: North-Holland. [132]

Elliott, G., and Timmermann, A. (2004), "Optimal Forecast Combinations Under General Loss Functions and Forecast Error Distributions," *Journal of Econometrics*, 122, 47–79. [133]

Hand, D. J., and Vinciotti, V. (2003), "Local versus Global Models for Classification Problems: Fitting Models Where It Matters," *The American Statistician*, 57, 124–131. [132]

Hansen, B. E. (2005), "Challenges for Econometric Model Selection," *Econometric Theory*, 21, 60–68. [132]

——— (2007), "Least Squares Model Averaging," *Econometrica*, 75, 1175–1189. [133]

——— (2008), "Least Squares Forecast Averaging," *Journal of Econometrics*, 146, 342–350. [133]

Heckman, J. J. (1984), "The χ^2 Goodness of Fit Statistic for Models With Parameters Estimated From Microdata," *Econometrica*, 52, 1543–1547. [132]

Hjort, N. L., and Claeskens, G. (2003), "Frequentist Model Average Estimators" (with discussion), *Journal of the American Statistical Association*, 98, 879–899. [132-135]

——— (2006), "Focused Information Criteria and Model Averaging for Cox's Hazard Regression Model," *Journal of the American Statistical Association*, 101, 1449–1464. [132,135]

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999), "Bayesian Model Averaging: A Tutorial," *Statistical Science*, 14, 382–417. [133]

Kao, C. (1986), "Variable Selection Problem in the Censored Regression Models," *Economics Letters*, 22, 353–357. [132]

Laitila, T. (1993), "A Pseudo-R^2 Measure for Limited and Quantitative Dependent Variable Models," *Journal of Econometrics*, 56, 341–356. [132,140]

Leeb, H., and Pötscher, B. M. (2008), "Can One Estimate the Unconditional Distribution of Post-Model-Selection Estimators?" *Econometric Theory*, 24, 338–376. [133]

Leung, G., and Barron, A. R. (2006), "Information Theory and Mixing Least-Squares Regressions," *IEEE Transactions on Information Theory*, 52, 3396–3410. [133]

Magnus, J. R., Powell, O., and Prüfer, P. (2010), "A Comparison of Two Model Averaging Techniques With an Application to Growth Empirics," *Journal of Econometrics*, 154, 139–153. [133,142]

Magnus, J. R., Wan, A. T. K., and Zhang, X. (2011), "Weighted Average Least Squares Estimation With Non-Spherical Disturbances and an Application to the Hong Kong Housing Market," *Computational Statistics & Data Analysis*, 55, 1331–1341. [133,142]

Mroz, T. (1987), "The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumption," *Econometrica*, 55, 765–799. [140]

Patton, A. J., and Timmermann, A. (2007), "Properties of Optimal Forecasts Under Asymmetric Loss and Nonlinearity," *Journal of Econometrics*, 140, 884–918. [133]

Pötscher, B. M. (2006), "The Distribution of Model Averaging Estimators and an Impossibility Result Regarding Its Estimation," *IMS Lecture Notes—Monograph Series: Time Series and Related Topics*, 52, 113–129. [141]

Schomaker, M., Wan, A. T. K., and Heumann, C. (2010), "Frequentist Model Averaging With Missing Observations," *Computational Statistics & Data Analysis*, 54, 3336–3347. [133]

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464. [135]

Tobin, J. (1958), "Estimation or Relationships for Limited Dependent Variables," *Econometrica*, 26, 24–36. [132]

Vaida, F., and Blanchard, S. (2005), "Conditional Akaike Information for Mixed-Effects Models," *Biometrika*, 92, 351–371. [132]

Varian, H. R. (1975), "A Bayesian Approach to Real Estate Assessment," in *Studies in Bayesian Econometrics and Statistics in Honor of Leonard J. Savage*, eds. S. E. Fienberg and A. Zellner, Amsterdam: North-Holland, pp. 195–208. [133]

Wan, A. T. K., and Zou, G. (2003), "Optimal Critical Values of Pre-Tests When Estimating the Regression Error Variance: Analytical Findings Under a General Loss Structure," *Journal of Econometrics*, 114, 165–196. [133]

Wan, A. T. K., Zhang, X., and Zou, G. (2010), "Least Squares Model Averaging by Mallows Criterion," *Journal of Econometrics*, 156, 277–283. [133]

Wan, A. T. K., Zou, G., and Ohtani, K. (2006), "Further Results on Optimal Critical Values of Pre-Test When Estimating the Regression Error Variance," *Econometrics Journal*, 9, 159–176. [133]

Yang, Y. (2001), "Adaptive Regression by Mixing," *Journal of the American Statistical Association*, 96, 574–588. [132,133]

Yuan, Z., and Yang, Y. (2005), "Combining Linear Regression Models: When and How?" *Journal of the American Statistical Association*, 100, 1202–1214. [132,133,136,138]

Zellner, A. (1986), "Bayesian Estimation and Prediction Using Asymmetric Loss Functions," *Journal of the American Statistical Association*, 81, 446–451. [133]

Zhang, X., and Liang, H. (2011), "Focused Information Criterion and Model Averaging for Generalized Additive Partial Linear Models," *The Annals of Statistics*, 39, 174–200. [133,135]