Regression Analysis with Many Specifications: A Bootstrap Method for Robust Inference

Peter Reinhard Hansen*

Brown University
Department of Economics, Box B
64 Waterman Street
Providence, RI 02912, USA
Email: Peter_Hansen@brown.edu

Preliminary Version. June, 2003

Abstract

We consider a set of linear regression models that differ in their choice of regressors, and derive a method for inference that properly controls for the set of models under investigation. The method is based around an estimate of the distribution for a class of statistics, which can depend on two or more models. An example is the largest R^2 over a set of regression models. The distribution will typically depend on all models in a complex way and does not have a known analytical form. This problem is solved by a bootstrap implementation that incorporates the relevant dependence across models.

We illustrate the method with an application where monthly stock returns are regressed on different subsets of lagged variables. The method is applied to estimate the distribution of the maximal R^2 under the null hypothesis of no explanatory power. In spite of the large number of models, the maximal R^2 is significant.

JEL Classification: C12, C2, C52, C53.

Keywords: Linear Regression; Data Mining; Model Comparison; Multiple Comparison; Bootstrap; R^2.

*A previous version of this paper carried the title "The Distribution of the Maximal R^2", and was presented at the NSF-NBER time-series conference in Pennsylvania, September 2002, the EC^2 conference, "Model Selection and Evaluation", Bologna, December 2002, and at Quebec City, May 2003. I thank Bruce Lehmann and Allan Timmermann for comments they made to this project while I was a graduate student at UCSD. I also thank Sean Campbell, Michael Jansson, Lutz Killian, and Gene Savin for comments and James Nason and Simon Potter for references. Financial support from the Danish Research Agency, grant no. 24-00-0363 and the Salomon Research Award at Brown University is gratefully acknowledged.

1 Introduction

Most textbook results about inference with regression models are based on a tacit assumption that the model is chosen independent of the data that are being analyzed. The reality of applied econometrics is often very different. The same data set is often used for both model selection (selection of the regressors) and the subsequent analysis, and in such a situation it is inappropriate to base inference on fixed-model results. Repeated use of data creates distortions, such as the bias caused by pre-testing and data mining. But repeated use of data is inevitable in econometrics analysis of observational data.[1]

In this paper, we derive a method for inference that can control for data mining and pre-testing in the context of regression models. We consider a situation where several regression models are estimated. These models differ in terms of their regressors but have the same dependent variable, and we seek the distribution of statistics that may be functions of statistics of the individual models, e.g. the maximal R^2 (across models) which is a function of the individual model's R^2s. We refer to statistics of this kind, as *inter-model statistics*.

The main contribution of this paper is a method that consistently estimates the distribution of a broad class of inter-model statistics. The method employs a bootstrap technique that circumvents the use of approximations, such as those based on probability inequalities, and we provide sufficient conditions that justify the use of the bootstrap in the context of regression models. We apply our method to derive the distribution of the maximal R^2 in a study, where monthly stock returns are regressed on lagged variables. The regression models include up to three regressors that are chosen from a set with 103 (mainly) macro-economic variables. Under the null hypothesis of no explanatory power, we estimate the distribution of the maximal R^2 (across regression models) and find strong evidence against the null, i.e., we conclude that certain regressors produce a significant R^2. It is comforting that the regressors that produce the largest R^2 are economically meaningful, as they measure changes in the level of economic

[1] As oppose to data collected from controlled experiments that may be repeated, to generate a new independent data set.

activities and the yield curve. It should be noted that our finding cannot be taken as evidence against the efficient markets hypothesis. The reason is that our results document a population relationship from an in-sample analysis, which need not imply that returns can be predicted out-of-sample.

Data mining and other types of data reuse are often criticized, but there need not be anything devious about it, as long as inference is adjusted accordingly. It is the lack of analytical results that makes it difficult to control for data reuse. In some cases, however, one can resort to approximative methods, such as those based on probability inequalities, e.g., the Bonferroni bound. The method that we introduce in this paper does not rely on bounds or ad hoc approximation and it applies quite generally. The bootstrap implementation of our method is similar to that applied to out-of-sample comparison of forecasting models, by White (2000) and Hansen (2001), who builds on results of Diebold and Mariano (1995) and West (1996). In spite of this similarity, the arguments that justify the use of the bootstrap in the present setting are quite different. The reason is that parameter uncertainty is the sole contributor to the uncertainties in the in-sample comparison of regression models, whereas this source often has negligible importance in out-of-sample comparisons of forecasting models, see West (1996). Whereas the aforementioned methods apply to out-of-sample evaluations our method applies to in-sample evaluation. Recent simulation studies by Inoue and Kilian (2002) indicate that the latter is more powerful, when the purpose is to access a population relationship, such as the significance of a parameter. Further, Inoue and Kilian (2002) call for the use of the bootstrap in practice, but do not derive theoretical results for the use of bootstrap methods in this context.

The method that we propose in this paper has a wide range of applications, and we discuss some of them next.

THE MAXIMAL R^2. Existing methods for approximating the distribution of the maximal R^2 (across several regression models) are discussed in Foster, Smith, and Whaley (1997). One method applies the Bonferroni bound which leads to a conservative method and a second method is known as the Rencher-Pun rule-of-thumb, (see Rencher and Pun (1980)), which

subsumes a particular covariance structure across regression models. Compared to these two approaches, the method proposed in this paper has two advantages. The first one is that the new method incorporates the sample information about the cross-model dependence, rather than assuming an extreme form of dependence (like the Bonferroni bound) or an ad hoc dependence structure (like the Rencher-Pun rule-of-thumb). The second advantage is that the new method does not rely on a restrictive distributional assumption, which are needed for the implementation of the two approximative methods. Our method can also be applied to estimate the bias in the R^2 that is caused by data mining, which can be used to create a more informative statistic in the context of data mining, e.g., the R^2 net of the estimated bias.

PRE-TESTING. The new method also applies to situations where pre-testing takes place. Model reduction by omitting insignificant regressors are known as the general-to-simple procedure. The estimators of a model that is selected in this manner, are likely to have properties different from those of a fixed model. In fact, Leeb and Pötscher (2003) have shown that the convergence in distribution of the estimators, cannot be uniform over the parameter space. Other interesting results have recently been derived by Danilov and Magnus (2001) in a classical setting with deterministic regressors and independent Gaussian distributed errors. Within this framework they are able to derive exact results that show the distortion that pre-testing creates. The method of this paper is capable of controlling for pre-testing in a general setting as it can consistently estimate the distribution of statistics of interest, where the analytical form of the distribution need not be known. For example, the unconditional distribution of a regression coefficient, which corresponds to a particular regressor (included in all models), can be estimated.

ROBUSTNESS. There is a large number of regression studies that apply economic growth rates as the dependent variable, and regressions of this kind are known as "growth regressions". The literature on growth regressions has questioned whether the significance of certain regressors are robust to the specification, e.g., would inclusion of additional regressors alter the result. Questions of this kind can also be addressed by the new method. Suppose that the regressors

are divided into two sets, one that contains regressors that are always included as regressors and one that contains regressors that are included in some, but not all models. Leamer (1983, 1985) referred to the first type of variables, as the *free variables* and to the second type of variables as the *doubtful variables*. We might be interested in evaluating whether the significance of a particular free variable is robust to the choice of included doubtful variables. Leamer proposed to estimate a confidence interval for the parameter in each of the m regression models, and define the "extreme bound" confidence interval to be the union of the m confidence intervals. Thus if zero lies within the extreme bound interval, the conclusion is that the significance is not robust. Levine and Renelt (1992) applied the extreme bound method to regression models using growth rates as the dependent variable, and did not find any "significant" regressors that satisfied this kind of robustness. Another way of describing the extreme bound method is that the method concludes that a significance is not robust, if the parameter is not significant is a single model. Since this conclusion is made without regard to the validity of this model, and hence the validity of the inference from this model, this method has obvious flaws. A series of papers, Sala-i-Martin (1997a, 1997b) and Doppelhofer, Miller, and Sala-i-Martin (2000) have criticized the extreme bound method for being too conservative, and proposed alternative methods. The method that we propose can also be applied in this context, as one could derive the distribution of the smallest t-statistic across models, and then evaluate whether the smallest t-statistic is significantly larger than this value. A better approach would be to take the smallest t-statistic over models that are consistent with the data, and not rejected in comparisons with other models.

MODEL SELECTION. Vuong (1989) considered the comparison of two non-nested models (possibly overlapping) in terms of their (quasi) likelihood ratio. In the context of regression models, the technique of this paper makes it possible to generalize Vuong's result to the comparison of more than two models. We discuss this issue in more details towards the end of Section 2.

The method can also be applied by a sceptical econometrician who believes certain re-

gressors should be included, but is unwilling to include additional regressors, unless there is evidence in favor of increasing the number of regressors. This can be done by extending the restricted model in various directions, e.g., by adding k_a regressors. Let δ_j be the parameters associated with the additional regressors, and consider all permutations of choosing k_a variables from a (finite) set of potential additional regressors. For each of the vectors for parameters, $\delta_1, \ldots, \delta_m$, consider the F-statistic, F_j, of the hypotheses $H_{0j} : \delta_j = 0$, $j = 1, \ldots m$. The new method can derive the distribution of, e.g., $\max_j F_j$ under the hypothesis, $H_0 : \delta_1 = \cdots = \delta_m = 0$, which can be used to construct a test of H_0. A procedure of this kind can be used as a simple-to-general procedure, and a general-to-simple method can be constructed in a similar way.

Although the method can be applied in situation where thousands of regression models are estimated, as we do in our application, the paper does not advocate this practice. One can even argue that extensive data mining of this kind defies the spirit of econometrics, because economic theory is not being incorporated in the modelling of the data. Nevertheless, in applied econometrics it is not uncommon that a data set is used to estimate a large number of regression models, perhaps by several econometricians. The applied econometrician who finds the "right" combination of regressors may appear to have "struck gold" and get the result published, although it may only be "fool's gold" – the result of data mining. This problem is particularly relevant in situations where the number of potential regressors is large relative to the number of observations of each variable. A situation that is not uncommon in the econometric analyses of macroeconomic variables.[2]

A Baysian approach to the selection of regressors was derived in Barbieri and Berger (2002). Their method requires one to specify priors over parameters and models, and derives the posterior probabilities over the set of models. In a second step, one assigns posterior probabilities to regressor i by adding up the posterior probabilities of the models that included variable i. This is done for all variables, and those that obtain a posterior probability greater than a half,

[2]Hundreds of macro economic variables are being recorded for the US alone. Most of these have been recorded for less than 60 years at a monthly, quarterly, or annual frequency.

are included in the final model.

Notation: We let $\text{diag}(A_1, \ldots, A_n)$ denote the block-diagonal matrix that has the matrices, $A_1, \ldots, A_n$ in the diagonal, and for $x \in \mathbb{R}$, we let $\text{int}(x)$ denote the integer part of x. For a symmetric positive semi-definite matrix, A, we let $A^{1/2}$ represent the symmetric semi-definite matrix that satisfies $A = A^{1/2}A^{1/2}$. We use subscript, j, to index models, $j = 1, \ldots, m$. For a matrix, A, with full column rank, we define $A_\perp$, to be some matrix with full column rank, that satisfies $\det(A, A_\perp) \neq 0$ and $A'_\perp A = 0$. The number of regressors in model j is denoted by k_j, and the total number of regressors is $K_m = \sum_{j=1}^{m} k_j$, where m is the number of models. The regressors are selected from a finite number of regressors, which is denoted by K. The probability measure that governs the raw data is denoted by P, and the probability measure that define the bootstrap resamples is represented by Q. Convergence in distribution and almost surely are denoted by $\xrightarrow{d}$ and $\xrightarrow{a.s.}$ respectively, where the convergence is with respect to the product measure $P \otimes Q$. (For quantities that do not involve bootstrap resamples this is equivalent to convergence with respect to P.)

This paper is organized as follow. In Section 2 we present the model and derive the new method that consistently estimates the distribution of statistics from a broad class of statistics. The distribution of the maximal R^2 is one example, which we discuss in more detail. The new method is based on bootstrap techniques, and we provide conditions that validate these techniques. Section 3 contains an empirical analysis of the predictability of monthly stock returns, and Section 4 contains concluding remarks. Proofs are given in the appendix.

2 Model and Assumptions

We consider a time-series $(Y_t, X_t')'$, $t = 1, \ldots, n$, where Y_t is a scalar and where X_t is a vector of variables. We consider regression models that have Y_t as the dependent variable and "some" of the variables in X_t as the explanatory variables. We have in mind a situation where the dimension of X_t is too large, relative to the sample size, such that it is impossible to obtain sensible results by including all variables as regressors. Instead, we consider m models, where

model j includes a few regressors, which we denote by X_{jt}, $j = 1, \ldots, m$. Let k_j be the number of variables in X_{jt}, and let $K_m \equiv \sum_{j=1}^{m} k_j$. For now, we leave the structure of the models unspecified, and do not require the number of regressors, k_j, to be the same in all models, however we take the number of models, m, as given. Statistics from the m models are combined into a single statistic, S, that we call an *inter-model statistic*. For the purpose of inference, we seek the (asymptotic) distribution of S, which we denote by F_S. The dependence across models, makes it intractable to derive analytical expressions for F_S, except in very simple situations. Fortunately, it is possible to estimate F_S using bootstrap techniques under certain regularity conditions, including the following assumption of Goncalves and de Jong (2003), which is modified to the present context.

Assumption 1 *Let $Z_t \equiv (Y_t, X_t')'$, $t = 1, \ldots, n$. The sequence $\{Z_t\}$ is stationary and ergodic, $E|Z_t|^{r+\epsilon} < \infty$ for $r > 2$ and some $\epsilon > 0$, and $\{Z_t\}$ is α-mixing of size $-r/(r-2)$.*

Before we turn to the formulation and estimation of regression models, we define the population quantities, $\mu_y \equiv E(Y_t)$ and $\sigma_y^2 \equiv \text{var}(Y_t)$, and for each set of regressors, X_{jt}, we define

$$\begin{aligned} \Sigma_{jj} &\equiv \text{var}(X_{jt}), \qquad \Sigma_{jy} \equiv \text{cov}(X_{jt}, Y_t), \qquad \mu_{x_j} \equiv E(X_{jt}), \\ \beta_j &\equiv \Sigma_{jj}^{-1}\Sigma_{jy}, \qquad \mu_j \equiv \mu_y - \beta_j'\mu_{x_j}, \qquad j = 1, \ldots, m. \end{aligned}$$

These quantities are well-defined under Assumption 1, and we consider the m regression models,

$$Y_t = \beta_j' X_{jt} + \mu_j + \varepsilon_{jt}, \qquad t = 1, \ldots, n, \quad j = 1, \ldots, m, \tag{1}$$

where X_{jt} is the vector of regressors that are used in model j. Note that the "errors", ε_{jt}, are defined by (1), and this construction ensures that $\{\varepsilon_{jt}\}$ inherits the stationarity from $(Y_t, X_t')'$ and that ε_{jt} is uncorrelated with X_{jt}. The regression parameters are estimated by least squares and it is convenient to define $\bar{Y} \equiv n^{-1}\sum_{t=1}^{n} Y_t$, $S_{yy} \equiv n^{-1}\sum_{t=1}^{n}(Y_t - \bar{Y})^2$, and for each of the

m models, we define

$$\bar{X}_j \equiv n^{-1}\sum_{t=1}^{n} X_{jt}, \quad S_{jj} \equiv n^{-1}\sum_{t=1}^{n} X_{jt}^c X_{jt}^{c\prime}, \quad S_{jy} \equiv n^{-1}\sum_{t=1}^{n} X_{jt}^c Y_t,$$

where $X_{jt}^c \equiv X_{jt} - \bar{X}_j$. It then follows that the least squares estimators are given by $\hat{\beta}_j \equiv S_{jj}^{-1} S_{jy}$ and $\hat{\mu}_j \equiv \bar{Y}_j - \hat{\beta}_j \bar{X}_j$, $j = 1, \ldots, m$. Next, we stack the regression parameters and their corresponding estimators into the vectors, $\boldsymbol{\beta}_0 \equiv (\beta_1', \ldots, \beta_m')'$ and $\widehat{\boldsymbol{\beta}} \equiv (\hat{\beta}_1', \ldots, \hat{\beta}_m')'$, and similarly we define

$$\boldsymbol{\psi}_0 \equiv (\sigma_y^2, \mu_y, \Sigma_{1y}, \Sigma_{11}, \mu_{x_1}, \ldots, \Sigma_{my}, \Sigma_{mm}, \mu_{x_m}),$$

and its sample analog $\widehat{\boldsymbol{\psi}} \equiv (S_{yy}, \bar{Y}, S_{1y}, S_{11}, \bar{X}_1, \ldots, S_{my}, S_{mm}, \bar{X}_m)$. The inter-model statistics that we consider in the following, will be function of $n^{1/2}\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\psi}}$. So the asymptotic distributions of an inter-model statistic is given by a transformation of the joint (across models) distribution of $n^{1/2}\widehat{\boldsymbol{\beta}}$, and the probability limit of $\widehat{\boldsymbol{\psi}}$. It is by using the joint distribution of $\widehat{\boldsymbol{\beta}}$, rather than combining those of $\hat{\beta}_1, \ldots, \hat{\beta}_m$, that we can achieve a consistent estimate of the distribution of an inter-model statistic.

Lemma 1 *Given Assumption 1 it holds that* $\widehat{\boldsymbol{\psi}} \xrightarrow{a.s.} \boldsymbol{\psi}_0$, *and*

$$n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} N(0, \boldsymbol{\Omega}),$$

where $\boldsymbol{\Omega} \equiv \boldsymbol{\Sigma}^{-1} V \boldsymbol{\Sigma}^{-1}$, $\boldsymbol{\Sigma} \equiv \mathrm{diag}(\Sigma_{11}, \ldots, \Sigma_{mm})$, $V = \Gamma_0 + \sum_{i=1}^{\infty}(\Gamma_i + \Gamma_i')$, *and where*

$$\Gamma_{t-s} \equiv E\begin{pmatrix} X_{1t}^c \varepsilon_{1t}\varepsilon_{1s} X_{1s}^{c\prime} & \cdots & X_{1t}^c \varepsilon_{1t}\varepsilon_{ms} X_{ms}^{c\prime} \\ \vdots & \ddots & \vdots \\ X_{mt}^c \varepsilon_{mt}\varepsilon_{1s} X_{1s}^{c\prime} & \cdots & X_{mt}^c \varepsilon_{mt}\varepsilon_{ms} X_{ms}^{c\prime} \end{pmatrix}.$$

The lemma implies that $(\bar{Y}, S_{yy}, S_{jy}, S_{jj}, \bar{X}_j) \xrightarrow{a.s.} (\mu_y, \sigma_y^2, \Sigma_{jy}, \Sigma_{jj}, \mu_j)$ and that $n^{1/2}(\hat{\beta}_j - \beta_j) \xrightarrow{d} N(0, \boldsymbol{\Omega}_j)$, where $\boldsymbol{\Omega}_j = \Sigma_{jj}^{-1} V_{jj} \Sigma_{jj}^{-1}$, $V_{jj} \equiv \Gamma_{jj,0} + \sum_{i=1}^{\infty} \Gamma_{jj,i} + \Gamma_{jj,i}'$, and $E\left[X_{j,t}^c \varepsilon_t \varepsilon_{t-i} X_{j,t-i}^{c\prime}\right]$

for $j = 1, \ldots, m$. So Lemma 1 contains the well-known result for a single model, as a special case. Lemma 1 could have been formulated in more general terms, by including the estimators for μ_j, $j = 1, \ldots, m$. However, the current formulation will suffice for our analysis and makes it possible to express certain limits in terms of the covariance matrices.

The usefulness of Lemma 1 is made clear below, where we consider transformations of the parameters. The transformation will have the form of the following assumption.

Assumption 2 *The mappings $r : \mathbb{R}^{K_m} \curvearrowright \mathbb{R}^p$ and $g : \mathbb{R}^p \times \mathbb{R}^q \curvearrowright \mathbb{R}^s$ are Borel measurable; r is differentiable with continuous derivative; and $g(y, z)$ is continuous at all points in $C \times \{\psi_0\}$, where $C \subset \mathbb{R}^p$ and satisfies $\lambda(\mathbb{R}^p \backslash C) = 0$, where λ is the Lebesgue measure.*

The assumption of measurability is needed in order to transform the asymptotic distribution of $\widehat{\boldsymbol{\beta}}$. Since r is continuous, the measurability of r is trivially satisfied, whereas the assumption is more substantive for the mapping g, which is allowed to be discontinuous.

Theorem 2 *Consider the random variables $\widehat{\boldsymbol{\xi}} \equiv g(n^{1/2} r(\widehat{\boldsymbol{\beta}}), \widehat{\boldsymbol{\psi}})$, and let $F_{\boldsymbol{\xi},n}$ be the cdf of $\widehat{\boldsymbol{\xi}}$. Suppose that Assumption 1 holds and that r and g satisfy Assumption 2. Let $F_{\boldsymbol{\xi}}$ denote the distribution of $\boldsymbol{\xi}_0 \equiv g(\tilde{Z}, \boldsymbol{\psi}_0)$, where $\tilde{Z} \sim N(0, \dot{r}(\boldsymbol{\beta})' \boldsymbol{\Omega} \dot{r}(\boldsymbol{\beta}))$, where $\dot{r}(\boldsymbol{\beta}) \equiv \left[\frac{\partial r_l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_{0,j}}\right]_{j=1,\ldots,K_m,\, l=1,\ldots,p}$. If $r(\boldsymbol{\beta}_0) = 0$, then $\lim_{n\to\infty} F_{\boldsymbol{\xi},n}(z) = F_{\boldsymbol{\xi}}(z)$, for all continuity points of $F_{\boldsymbol{\xi}}$.*

The mapping r is used to impose restrictions on the parameters, $r(\boldsymbol{\beta}_0) = 0$, and in most cases we consider restriction on the form $r(\boldsymbol{\beta}_0) = A'\boldsymbol{\beta}_0 - a = 0$, where $A = \dot{r}(\boldsymbol{\beta})$ (for all $\boldsymbol{\beta}$) is a known matrix and a is a known vector. Without loss of generality we assume that A has full column rank, which leads to the equivalent formulation of the restrictions, given by $\boldsymbol{\beta}_0 = H\boldsymbol{\varphi}_0 + h$, where $H = A_\perp$ and where $h = A\left(A'A\right)^{-1} a$.

This result of Theorem 2 enables us to obtain the asymptotic distribution for a class of inter-model statistics. We illustrate this by the following example.

Example 1 *Consider the average R^2, given by $R^2_{\text{ave}} = m^{-1} \sum_{j=1}^m R_j^2$. Under the hypothesis*

that $\beta_1 = \cdots = \beta_m = 0$, *we set* $r(x) = x$ *and let* g *be given by*

$$g(n^{1/2} r(\widehat{\boldsymbol{\beta}}), \widehat{\boldsymbol{\psi}}) = \frac{n}{m} \frac{\widehat{\boldsymbol{\beta}}' \operatorname{diag}(S_{11}, \ldots, S_{mm}) \widehat{\boldsymbol{\beta}}}{S_{yy}},$$

which is continuous (*assuming* $\sigma_y^2 > 0$). *Since*

$$n R^2_{\text{ave}} = n\, m^{-1} \left(\frac{\hat{\beta}_1' S_{11} \hat{\beta}_1}{S_{yy}} + \cdots + \frac{\hat{\beta}_m' S_{mm} \hat{\beta}_m}{S_{yy}} \right) = g(n^{1/2} r(\widehat{\boldsymbol{\beta}}), \widehat{\boldsymbol{\psi}}),$$

it follows by Theorem 2, that the asymptotic distribution of $n R^2_{\text{ave}}$ *is given by*

$$\frac{1}{\sigma^2} Z_I' \left(\Omega^{1/2} \Sigma \Omega^{1/2} \right) Z_I = \frac{1}{\sigma^2} Z_I' \left(\Sigma^{-1/2} V \Sigma^{-1/2} \right) Z_I,$$

where $Z \sim N_{K_m}(0, I)$, *since* $S_{yy} \xrightarrow{a.s.} \sigma^2$ *and* $\operatorname{diag}(S_{11}, \ldots, S_{mm}) \xrightarrow{a.s.} \Sigma$, *under the null hypothesis*.

The distribution in the previous example, $Z' \left(\Sigma^{-1/2} V \Sigma^{-1/2} \right) Z$, is recognized as a linear combination of independent χ^2-distribution variables, where the weights depend on the eigenvalues of $\left(\Sigma^{-1/2} V \Sigma^{-1/2} \right)$. So it is possible to derive an analytical expression for the asymptotic distribution of $n R^2_{\text{ave}}$, however this is not the case for most inter-model statistics, and in these situations the bootstrap implementation is particularly useful.

As stated in the introduction, we follow the terminology of Leamer (1983, 1985) and call the variables, which are included in all models, the *free variables*, and the remaining variables of X_t are called the *doubtful variables*. To simplify matters, we assume that the number of doubtful variables is the same across models, we let k be the number of regressors, and denote the number of doubtful regressors by q. We divide the parameters into $\beta_j = (\gamma_j', \delta_j')'$, where γ_j are the parameters that correspond to the free variables and δ_j are the parameters that correspond to the doubtful variables in model j. Note that all models are identical if $\delta_1 = \cdots = \delta_m = 0$, in which case we define $\gamma \equiv \gamma_1 = \cdots = \gamma_m$.

We shall consider hypotheses that take the form, $H_0 : \delta_1 = \cdots = \delta_m = 0$ and $r_1(\gamma) = 0$. So we consider hypotheses, where all relevant regressors are in the set of free variables. By testing the zero-restrictions on δ_j, $j = 1, \ldots, m$, we are able to evaluate whether there is evidence that any additional regressor are needed. If it is possible to estimate the comprehensive model that includes all variables as regressors, we can test this hypthesis with an F-test. However, we have in mind a situation where the number of variables is so large, that there is not sufficient data to estimate the full model. Instead we consider models that only contains a limited number of regressors, and compare the restricted model to several alternative models that includes a few additional doubtful variables. This leads to a multiple comparisons problem, and the inference in such problems are often quite involved. However, we circumvent the complications by using a bootstrap implementation that makes it simple to test the hypothesis.

2.1 Bootstrap Implementation

Our implementation is based on the stationary bootstrap (SB) of Politis and Romano (1994) as this choice facilitates the analysis.[3] Our assumptions that justify the use of the bootstrap implementation is based on the relaxed (moment) conditions due to Goncalves and de Jong (2003). The bootstrap serves as a tool to take the dependence across models into account, and our implementation is similar to that used to compare forecasting models by White (2000) and Hansen (2001). It should be noted that there is an important difference between the analysis that justifies the use of the bootstrap in the present setting and that of comparing forecasting models. White (2000) showed the consistency of the stationary bootstrap in the context of out-of-sample comparison of forecasting models. These results do not apply to our analysis, which is concerned with in-sample comparison. In the in-sample analysis the major "noise-component" arises from the estimation of parameters. This is in sharp contrast to the out-of-sample analysis, where the contributions from the parameter estimators, is only one of two components, and in fact, the standard assumptions made in comparisons of forecasting models, ensure that the contribution

[3] Nevertheless, we postulate that the results can be obtained with an implementation based on the block-bootstrap of Kunsch (1989).

from estimation uncertainty vanishes asymptotically, see West (1996). In the context of out-of-sample comparison and testing a composite hypothesis, Hansen (2001, 2003) has shown some unfortunate properties of this bootstrap implementation, and suggested modifications to correct these problems. The hypotheses that we consider below do not have a composite form that requires these modifications.

The resamples of the stationary bootstrap are defined as follows.

Definition 1 (Stationary Bootstrap) *Let $\rho \in (0, 1]$, (the dependence parameter). A stationary bootstrap re-sample, Z_t^*, is generated from the sample, $Z_1, \ldots, Z_n$, by defining $Z_t^* \equiv Z_{\tau_{(t)}}$, where $\tau_{(t)}$, $t = 1, \ldots, n$, is generated as follows.*

Let $\upsilon_1, \ldots, \upsilon_n, \eta_1, \ldots, \eta_n$ be independent and uniformly distributed on $[0, 1)$, and let $\tau_{(1)} = 1 + \text{int}(\upsilon_1 n)$, and for $t = 2, \ldots, n$, let

$$\tau_{(t)} = 1 + \text{int}(\upsilon_t n)1_{\{\eta_t \leq \rho\}} + \tau_{(t-1)}1_{\{\tau_{(t-1)} < n\}}1_{\{\eta_t > \rho\}},$$

where $1_{\{\cdot\}}$ is the indicator function.

Thus, a re-sample, $\{Z_t^*\}$, is composed of blocks of the original sample, $\{Z_t\}$, where the block-length is geometrically distributed, and where the expected length of each block is ρ^{-1}. A property of this construction, is that a re-sample, Z_t^*, is stationary and ergodic, conditional on $Z_1, \ldots, Z_n$.[4] In our asymptotic analysis, we let ρ depend on the sample size, and write ρ_n. Specifically, we let $\rho_n \to 0$ (at a suitable rate) to ensure that the resamples (eventually) will resemble the time-dependence of the observed process.

We let $\{(Y_t^*, X_t^{*\prime})'\}$ denote a SB re-sample of our observations $\{(Y_t, X_t')'\}$ and calculate the statistics, S_{yy}^*, $\bar{Y}^*$, S_{jy}^*, S_{jj}^*, and $\bar{X}_j^*$, where the statistics are defined by the formulae for S_{yy}, $\bar{Y}$, S_{jy}, S_{jj}, and $\bar{X}_j$, after having replaced (Y_t, X_{jt}) by (Y_t^*, X_{jt}^*). It follows that $\hat{\beta}_j^* = S_{jj}^{*-1} S_{jy}^*$ is the least squares estimator for β_j when regressing Y_t^* on X_{jt}^* (and a constant). Similarly, we

[4] This follows directly from the theory of Markov chains.

define

$$\widehat{\psi}^* = (S^*_{yy}, \bar{Y}^*, S^*_{1y}, S^*_{11}, \bar{X}^*_1, \ldots, S^*_{my}, S^*_{mm}, \bar{X}^*_m).$$

Lemma 3 *If $(Y_t, X_t')'$ has finite second moment, then $\widehat{\psi}^* \xrightarrow{a.s.} \psi_0$.*

Assumption 3 *Let $\mathbf{Z}_t \equiv (\varepsilon_{1,t}X'_{1,t}, \ldots, \varepsilon_{m,t}X'_{m,t})'$. For some $r > 2$ and some $\epsilon > 0$, it holds that $E|\mathbf{Z}_t|^{r+\epsilon} < \infty$ and that $\{\mathbf{Z}_t\}$ is α-mixing of size $-r/(r-2)$.*

Given our definition of $\{\varepsilon_{j,t}\}$, $j = 1, \ldots, m$, it is clear that assumptions on $\{Y_t, X_t'\}$ have implications for $\{\varepsilon_{j,t}\}$. The following lemma reassures us that Assumption 3 does not violate our previous assumptions on $\{Y_t, X_t'\}$.

Lemma 4 *Assumption 3 is implied by Assumption 1, strengthened by $E\left|(Y_t, X_t')'\right|^{2(r+\epsilon)} < \infty$.*

Theorem 5 *Suppose that $\rho_n \to 0$ and $n\rho_n^2 \to \infty$ as $n \to \infty$. Given Assumption 3, it holds for all $\epsilon > 0$ that*

$$\sup_z \left| P^*(n^{1/2}(\widehat{\beta}^* - \widehat{\beta}) \leq z) - P(n^{1/2}(\widehat{\beta} - \beta_0) \leq z) \right| \xrightarrow{p} 0. \tag{2}$$

Let $F^*_{\xi,n}$ denote the cdf of $\widehat{\xi}^* \equiv g(n^{1/2}r(\widehat{\beta}^*) - n^{1/2}r(\widehat{\beta}), \widehat{\psi}^*)$, where r and g are as in Assumption 2, then, $F^*_{\xi,n}(z) \to F_\xi(z)$, for all the continuity points of F_ξ.

COMMENT 1: The probability P^* is evaluated conditional on the sample $X_1, \ldots, X_n$, and the result states that the bootstrap distribution is close to the true distribution, in the sense of (2).

COMMENT 2: Theorem 5 applies to least squares estimation of regression models. But the result can be generalized to a broader class of estimators, that includes quasi maximum likelihood estimators (QMLEs) and generalized method of moments (GMM) estimators under appropriate assumptions, although we shall not pursue this here. For the case with $m = 1$, (2) has been established by Goncalves and White (2003) for the aforementioned class of estimators.

2.2 Applications

The following corollary and examples show how Theorem 5 can be applied to obtain bootstrap estimates of quantities that are functions of the R^2s, F-statistics, or t-statistics. The implementation requires that we impose the null hypothesis on the bootstrap variables. However, it is not streight-forward to impose a population R^2 to have a particualar value with a non-parametric bootstrap. Instead we make use of a one-to-one relation between the R^2 and a particular F-statistic.[5] The well-known relation is

$$R_j^2 = \frac{F_j}{F_j + (n-k-1)/k},$$

where F_j is the F-statistic of the hypothesis $\beta_j = 0$, $j = 1, \ldots, m$. Thus, for each resample, we calculate $(F_{1,b}^*, \ldots, F_{m,b}^*)$, where $F_{j,b}^*$ is the F-statistic of the hypothesis $\beta_j = \hat{\beta}_j$, based on the bth re-sample, $b = 1, \ldots, B$, and define

$$R_{j,b}^{*2} \equiv \frac{F_{j,b}^*}{F_{j,b}^* + (n-k-1)/k}.$$

Note that $R_{j,b}^{*2}$ is not the R^2 of the bth resample using model j's regressors, it is derived from the F-statistic, that ensures that the bootstrap distribution approximates the distribution of R_j^2, under the appropriate null hypothesis.

Corollary 6 (Bootstrap Distribution of the Maximal R^2) *Consider the maximal R^2 across models, $R_{\max}^2 = \max_{j=1,\ldots,m} R_j^2$. Under the hypothesis that $\beta_1 = \cdots = \beta_m = 0$, the distribution of $R_{\max}^2$, $F(z) = P\left(R_{\max}^2 \leq z\right)$ can be approximated by*

$$\hat{F}(z) = \frac{1}{B}\sum_{b=1}^{B} 1_{\left\{R_{\max,b}^{*2} \leq z\right\}}, \quad 0 \leq z \leq 1, \tag{3}$$

for B and n sufficiently large, where $R_{\max,b}^{*2} \equiv \max_{j=1,\ldots,m} R_{j,b}^{*2}$.

[5] When resampling the variables $(Y_t, X_{j,t})$, the population parameters of the resample are given by the estimated parameters in model j, $\hat{\beta}_j$. Other values of the parameters would generate the same value of the R^2, but those are irrelevant for the purpose of resampling.

Example 2 (The Maximal R^2 under a Non-Trival Null Hypothesis) *Suppose instead that we are interested to measure how much the R^2 can be increased by adding a fixed number of additional regressors. Thus the object of interest is the distribution of $S = R^2_{\max} - R^2_0$, under the null hypothesis, $H'_0 : \delta_j = \cdots = \delta_j = 0$, where R^2_0 is the R^2 of the model that only includes the p free variables as regressors. We proceed by estimating the restricted model (with the p free variables) for the original sample to obtain the estimator $\hat{\gamma}$. For each of the resamples we also estimated the restricted model, and define $R^{*2}_{0,b} = F^*_{0,b}/[F^*_{0,b} + (n-p-1)/p]$, where $F^*_{0,b}$ is the F-statistic for the hypothesis that $\gamma = \hat{\gamma}$. By defining $R^{*2}_{\max,b}$ as above, the distribution of S, can now be estimated by $\hat{F}(z) = \frac{1}{B}\sum_{b=1}^{B} 1_{\left\{R^{*2}_{\max,b} - R^*_{0,b} \leq z\right\}}$. Thus the expected boost in the R^2 from searcing over the m larger models is estimated by $\frac{1}{B}\sum_{b=1}^{B} R^{*2}_{\max,b} - R^*_{0,b}$.*

The distribution of F-statistics and t-statistics can be estimated in a similar way.

Example 3 (Pre-Testing) *Suppose that we have two possible regressors, X_{1t} and X_{2t}, and we estimate the regression models:*

$$\begin{aligned} Y_t &= \gamma_1 X_{1t} + \mu_1 + \varepsilon_{1t}, \\ Y_t &= \gamma_2 X_{1t} + \delta_2 X_{2t} + \mu_2 + \varepsilon_{2t}. \end{aligned}$$

If δ_2 is found to be significant in the second regression equation, then $\hat{\gamma}_2$ is reported otherwise $\hat{\gamma}_1$ is reported. The procedure can be represented by the mapping g, which is defined by

$$g(\widehat{\beta}, \widehat{\psi}) = \begin{cases} \hat{\gamma}_1 & \text{if } \frac{|\hat{\delta}_2|}{\sigma_{\hat{\delta}_2}} \leq c_r, \\ \hat{\gamma}_2 & \text{otherwise,} \end{cases}$$

where $|\hat{\delta}_2|/\sigma_{\hat{\delta}_2}$ is the absolute value of the t-statistic of $\delta_2 = 0$. We seek the distribution of the estimator, $\hat{\gamma}$, which this procedure leads to, under the hypothesis, $H_0 : \gamma_1 = \gamma_2 = 0$. Dependent on the t-statistic, the procedure will report either $\hat{\gamma}_1$ or $\hat{\gamma}_2$, so the distribtion of $\hat{\gamma}$, is a complicated mixture of the distribution of $\hat{\gamma}_1$ and $\hat{\gamma}_2$. We can express $\hat{\gamma}$ as the mapping, $\hat{\gamma} = g(\widehat{\beta}, \widehat{\psi})$. We note that this mapping is dicontinuous, and the tresshold where we change

from reporting $\hat{\gamma}_1$ *to* $\hat{\gamma}_2$. *Now if* $\delta_2 \neq 0$ *then* $\frac{|\hat{\delta}_2|}{\sigma_{\hat{\delta}_2}} \to \infty$ *as* $n \to \infty$, *in which case* $\hat{\gamma}_2$ *is always reported. On the other hand, if* $\delta_2 = 0$, *then the set of discontinuity points has measure zero. So g satisfies Assumption 2, and Theorem 5 can be applied to estimate the distribution of* $\hat{\gamma}$.

Example 4 (Vuong Type Likelihood Ratio Test) *If we assume a Gaussian likelihood for the regression models and let* $L_{\max,j}$ *be the maximum value of model j's likelihood, then we can consider* $(\log L_{\max,1}, \ldots, \log L_{\max,m})$. *Hypotheses, such as* $E(n^{-1} \log L_{\max,i}) = E(n^{-1} \log L_{\max,j})$ *for all* $i, j = 1, \ldots, m$, *or* $E(n^{-1} \log L_{\max,1}) \geq E(n^{-1} \log L_{\max,j})$ *for all* $j = 2, \ldots, m$, *can be tested using test statistics that are functions of* $(\log L_{\max,1}, \ldots, \log L_{\max,m})$, *which are themselves functions of* $\widehat{\beta}$ *and* $\widehat{\psi}$.

3 Empirical Analysis of Monthly Stock Returns

In this section, we estimate a large number of regression models, where monthly stock returns are regressed on lagged variables. The application illustrates how the new method can be applied to control for data mining. In this application we need to a control for the large number of regression models that we estimate, prior to reporting the largest R^2. For additional discussion on data mining in financial econometrics, see Merton (1987), Ross (1989), and Lo and MacKinlay (1990).

There are several studies that evaluate the predictability of stock and bond returns using regression models, and measure the predictability by the R^2, see e.g., Campbell and Shiller (1988) and Fama and French (1989). The R^2 is often found to be surprisingly large in "long-horizon predictions", and the significance of this observations have been discussed in the literature. Hodrick (1992) and Kirby (1997) note that autocorrelation distorts the size of tests in long-horizon forecasting environments, and argue that the R^2 may be spuriously large in long-horizon predictions, whereas Patelis (1997) argues that monetary policy creates the predictability. Another possibility is that the predictability have spuriously have been discovered by the research community after having explored a large number of possible regression models. Our new method, is capable of controlling for both the autocorrelation and the mining over a set of regressor

models.

We consider regression models of the form of (1), $Y_t = \beta_j' X_{jt} + \mu_j + \varepsilon_{jt}$, $t = 1, \ldots, n$, $j = 1, \ldots, m$, where the dependent variable, Y_t, is the monthly return on the Dow Jones Industrial Average (DJIA) index. The regressors of model j, X_{jt}, are predetermined variables, primarily macro economic variables. We also consider the regression, where the dependent variable, Y_t, is the monthly excess return over the return on a one-month Treasury bill. The subset of regressors, X_{jt}, are selected from X_t, which is a vector of $K = 103$ variables. These variables include: lagged returns; interest rates; monetary variables; business cycle indicators; unemployment variables; price indexes; and production and consumption data. Acknowledging that several macro variables are published with a lag and therefore not observed in real time, we have lagged some of the variables by up to three periods to make it more realistic that our regressors were observed at time $t - 1$. Our data are monthly observation that span the period 1959:01 to 1998:01. The effective sample period (which is available for estimation) is reduced to 1960:07 – 1998:01, because some regressors are constructed from variables that are lagged by up to 18 months. The data sources are: Federal Reserve Economic Data (FRED), Domestic Economic Data (DRI), and the Center for Research in Security Prices (CRSP), and the variables are listed in Table 1.

All the variables in X_t are classified as doubtful variables and we consider all distinct models that include $k = 1$, 2, or 3 regressors, (in addition to a constant). This result in $m = \binom{K}{k}$ models. Thus for $k = 1$ we have $m_1 = 103$ models, for $k = 2$ we have $m_2 = 5,253$ models, and for $k = 3$ we have $m_3 = 176,851$ models.

We consider the maximal R^2 under the hypothesis that none of the variables have any explanatory power, $H_0 : \beta_0 = (\beta_1', \ldots, \beta_m')' = 0$. Even if the null hypothesis is correct, some of the models are likely to produce an R^2 that would appear to be significant, if we did not adjust the critical value for the (large) number of models.

The results are presented in Table 2, which displays convincing evidence against the null hypothesis. Even amongst the models with a single regressor we find significant regressors, and

the models with two and three regressors provide even stronger evidence against the hypothesis. Also, the findings are robust to the choice of dependent variable (returns or excess returns). All results are based on $B = 1,000$ bootstrap replications where $q = \frac{1}{2}$ was used as the dependence parameter.

The "Miner (Beta)" critical value is calculated from the Beta-distribution, $\mathcal{B}\left(\frac{k}{2}, \frac{n-k-1}{2}\right)$. This would be the appropriate distribution to use, if the R^2 had been calculated from a single model with deterministic regressors and innovations that were iid Gaussian distributed. The "BF (Beta)" employs the Beta-distribution and the Bonferroni inequality to derive critical values. It is interesting to note that "Bootstrap" critical values can exceed those of "BF (Beta)". For example, for $k = 1$, the "Bootstrap" critical value exceeds that of the Bonferroni bound. This may appear surprising since the latter is called a 'bound' for the critical value. However, the Bonferroni bound method is based on Beta-distribution that is justified by assumptions that are known to be invalid in the present context. A kernel estimation of the bootstrap-density for the case $k = 2$ (where the dependent variable is excess returns) is shown in Figure 1.

Given the significance of the maximal R^2 it is interesting to analyze which of the variables that led to to a large R^2, and Table 3 presents the regressors that resulted in the maximal R^2, for the two dependent variables, returns and excess returns, and for models with one, two, and three regressors.

From Table 3 it is comforting to see that the regressors with the best sample prediction are key economic variables, rather than more obscure variables. The yield on the ten-year treasury bond is selected in all but one model, and the second most important variable is DMANEM_2, which is defined as the change in employment in the manufacturing sector (lagged two month). So an indicator of economic activity, and a measure of the yield curve turn out to be the best predictors of returns.

The in-sample prediction errors using the best three-variable model to predict excess returns are plotted in Figure 2.

There is some evidence that the predictability carries over to out-of-sample forecasts. Pe-

saran and Timmermann (1995) found that in a real-time setting with recursively estimated parameters, and various model selection criteria, the selected models would historically have produced valuable forecasts, especially during periods when interest rates were relative volatile. Although, it should be added that their analysis did not fully account for the 'mining' over models.

4 Conclusion

We have introduced a new method for robust inference in the situation where several regression models are estimated. We considered a general class of statistics that we referred to as inter-model statistics, and we have showed how the distribution of an inter-model statistic, can be estimated under fairly general assumptions. In the general situation, there does not exist an alternative method, and in the specific cases where alternative methods exist, the latter are inferior, as they neglect the sample information about cross-model dependence, which is important for the analysis. Our method implicitly estimates the cross-model dependence and exploits this to produce a consistent estimate of the relevant distribution. For this reason we believe that our method is an important contribution to the literature.

The applicability was illustrated with an in-sample regression study, where monthly returns on the Dow Jones industrial average index were regressed on lagged variables. The results showed that the largest R^2 is significant, even in the set of models that only include a single regressor, and the evidence is stronger as the number of regressors is increased. This finding cannot be taken as evidence against the efficient market hypothesis, because full sample estimates are not available for in-sample trading strategies, and an out-of-sample analysis would presumably lead to a different result.

Extensions and future research include empirical applications of the method, and comparisons to approximative methods in the situation where such apply. Also, a much deeper analysis of the extension to multiple comparisons of likelihoods is needed – in a more general framework than the regression models considered in this paper.

A Appendix of Proofs

Proof of Lemma 1. Given Assumption 1, the strong consistency of $\widehat{\psi}$ follows by the ergodic theorem (the strong law of large numbers for stationary and ergodic precesses), and the asymptotic normality follows from

$$n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \begin{pmatrix} S_{11}^{-1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & S_{mm}^{-1} \end{pmatrix} n^{1/2} \sum_{t=1}^{n} \begin{pmatrix} X_{1t}^{c} \boldsymbol{\varepsilon}_{1t} \\ \vdots \\ X_{mt}^{c} \boldsymbol{\varepsilon}_{mt} \end{pmatrix},$$

since the first term converges in probability to Σ^{-1} and the second term converges in distribution to $N_{K_m}(0, V)$. ■

Proof of Theorem 2. A Taylor expansion of r about $\boldsymbol{\beta}_0$ yields

$$r(\widehat{\boldsymbol{\beta}}) = r(\boldsymbol{\beta}_0) + \dot{r}(\boldsymbol{\beta}_{0*})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0),$$

where $\boldsymbol{\beta}_{0*}$ lies between $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$. Since $\widehat{\boldsymbol{\beta}} \xrightarrow{a.s.} \boldsymbol{\beta}_0$, also $\boldsymbol{\beta}_{0*} \xrightarrow{a.s.} \boldsymbol{\beta}_0$, and by the continuity of $\dot{r}$ it follows that also $\dot{r}(\boldsymbol{\beta}_{0*}) \xrightarrow{a.s.} \dot{r}(\boldsymbol{\beta}_0)$. We have established that $n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} Z$, for some random variable, $Z \sim N(0, \boldsymbol{\Omega})$. So if we define $\tilde{Z} = \dot{r}(\boldsymbol{\beta}_0)Z$, it follows that $n^{1/2}r(\widehat{\boldsymbol{\beta}}) \xrightarrow{d} \tilde{Z}$, under the restriction, $r(\boldsymbol{\beta}_0) = 0$. Finally, since $(n^{1/2}r(\widehat{\boldsymbol{\beta}}), \widehat{\psi}) \xrightarrow{d} (\tilde{Z}, \psi_0)$ it follows by the continuous mapping theorem (for functions that are continuous almost everywhere w.r.t. the limit distribution of $(n^{1/2}r(\widehat{\boldsymbol{\beta}}), \widehat{\psi})$) that $\widehat{\boldsymbol{\xi}}_n \xrightarrow{d} \boldsymbol{\xi}_0$. ■

Lemma 7 *Let $\{Z_t\}$ be a stationary and ergodic sequence with finite first moment and define $\mu \equiv E(Z_1)$. Then*

$$\bar{Z}_n^* \xrightarrow{a.s.} \mu,$$

where $\bar{Z}_n^* = n^{-1}\sum_{t=1}^{n} Z_t^*$ and where Z_t^* is bootstrap re-sample as define in Definition 1.

Proof. We let $(\Omega, \mathcal{F}, P)$ denote the probability space that governs $\{Z_t\}$ and let $(\Lambda, \mathcal{G}, Q)$ be the probability space that governs $\{(\upsilon_t, \eta_t)\}$, i.e., the probability measure that governs the

bootstrap resamples. Then by the specification of (υ_t, η_t) (see Definition 1), it holds that $\{Z_t\}$, $\{\upsilon_t\}$, and $\{\eta_t\}$ are mutually independent. By the ergodic theorem we have that $\bar{Z}_n \xrightarrow{a.s.} \mu$, where $\bar{Z}_n = n^{-1}\sum_{t=1}^{n} Z_t$, so there exists a $F \subset \mathcal{F}$ with $P(F) = 1$, such that for any $\omega \in F$ and any $\epsilon > 0$, there exists an $N_{1\epsilon}(\omega)$, such that $|\bar{Z}_n(\omega) - \mu| < \epsilon/2$ for all $n > N_{1\epsilon}(\omega)$. Similarly, since $\{Z_t^*\}$ is stationary and ergodic (conditional on $Z_1, \ldots, Z_n$), there exists $G \in \mathcal{G}$ with $P(G) = 1$, such that for any $\lambda \in G$, any $\omega \in \mathbf{\Omega}$, and any $\epsilon > 0$, there exists an $N_{2\epsilon}(\lambda, \omega)$, such that $|\bar{Z}_m^*(\lambda, \omega) - \bar{Z}_n(\omega)| < \epsilon/2$ for all $m > N_{2\epsilon}(\lambda, \omega)$. So we have that for $(\omega, \lambda) \in F \times G$ and any $\epsilon > 0$, there exist an $N(\lambda, \omega) = \max(N_{1\epsilon}(\omega), N_{2\epsilon}(\lambda, \omega))$ such that $|\bar{Z}_n^*(\lambda, \omega) - \mu| \leq |\bar{Z}_n^*(\lambda, \omega) - \bar{Z}_n(\omega)| + |\bar{Z}_n(\omega) - \mu| < \epsilon$ for all $n > N(\lambda, \omega)$. This shows the strong consistency. ■

Proof of Lemma 3. The elements of $\widehat{\psi}^*$ are all linear combinations of sample averages of the form $\bar{Z}_n^* = n^{-1}\sum_{t=1}^{n} Z_t^*$, where $Z_t^* = X_t^*$, Y_t^*, $X_t^* X_t^{*\prime}$, Y_t^{*2}, or $X_t^* Y_t^*$. Lemma 7 shows that each of these, converges almost surely provided that $E(|(Y_t, X_t')'|^2) < \infty$. ■

Proof of Lemma 4. The identity $X_{j,t}\varepsilon_{j,t} = X_{j,t}Y_t - X_{j,t}X_{j,t}'\beta_j$, shows that $X_{j,t}\varepsilon_{j,t}$ inherits the mixing properties from $(Y_t, X_t')'$, which are given Assumption 1. The strengthened moment condition implies that $E|X_t Y_t|^{r+\epsilon} < \infty$ and $E|X_t X_t'|^{r+\epsilon} < \infty$ $j = 1, \ldots, m$, and by Minkowski's inequality:

$$\begin{aligned}
(E\left|X_{j,t}\varepsilon_{j,t}\right|^{r+\epsilon})^{\frac{1}{r+\epsilon}} &= (E\left|X_{j,t}Y_t + (-X_{j,t}X_{j,t}'\beta_j)\right|^{r+\epsilon})^{\frac{1}{r+\epsilon}} \\
&\leq (E\left|X_{j,t}Y_t\right|^{r+\epsilon})^{\frac{1}{r+\epsilon}} + (E\left|X_{j,t}X_{j,t}'\right|^{r+\epsilon})^{\frac{1}{r+\epsilon}}\left|\beta_j\right|,
\end{aligned}$$

it follows that $E\left|X_{j,t}\varepsilon_{j,t}\right|^{r+\epsilon} < \infty$, $j = 1, \ldots, m$. ■

Proof of Theorem 5. From the identity

$$S_{jy}^* = n^{-1}\sum_{t=1}^{n} X_{j,\tau_t}Y_{\tau_t} = n^{-1}\sum_{t=1}^{n} X_{\tau_t}(X_{j,\tau_t}'\beta_j + \varepsilon_{j,\tau_t}) = S_{jj}^*\beta_j + n^{-1}\sum_{t=1}^{n} X_{j,\tau_t}\varepsilon_{j,\tau_t},$$

we see that $n^{1/2}(\hat{\beta}_j^* - \beta_j) = S_{jj}^{*-1}n^{-1/2}\sum_{t=1}^{n} X_{j,\tau_t}\varepsilon_{j,\tau_t}$ and similarly

$$n^{1/2}(\hat{\beta}_j - \beta_j) = S_{jj}^{-1}n^{-1/2}\sum_{t=1}^{n} X_{j,t}\varepsilon_{j,t}.$$

Define $\mathbf{Z}_t \equiv (\varepsilon_{1,t}X'_{1,t}, \ldots, \varepsilon_{m,t}X'_{m,t})'$, $\mathbf{S}^* = \text{diag}(S^*_{11}, \ldots, S^*_{mm})$, and $\mathbf{S} = \text{diag}(S_{11}, \ldots, S_{mm})$. From the identity $(\widehat{\boldsymbol{\beta}}^* - \widehat{\boldsymbol{\beta}}) = (\widehat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}_0) - (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ it follows that

$$\begin{aligned} n^{1/2}(\widehat{\boldsymbol{\beta}}^* - \widehat{\boldsymbol{\beta}}) &= n^{1/2}(\widehat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}_0) - n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\ &= \mathbf{S}^{*-1}n^{-1/2}\sum_{t=1}^{n}\mathbf{Z}_{\tau_t} - \mathbf{S}^{-1}n^{-1/2}\sum_{t=1}^{n}\mathbf{Z}_t \\ &= \mathbf{S}^{*-1}n^{-1/2}\sum_{t=1}^{n}(\mathbf{Z}_{\tau_t} - \mathbf{Z}_t) + (\mathbf{S}^{*-1} - \mathbf{S}^{-1})n^{-1/2}\sum_{t=1}^{n}\mathbf{Z}_t, \end{aligned}$$

where the first term equals $\mathbf{S}^{*-1}n^{-1/2}\sum_{t=1}^{n}(\mathbf{Z}_{\tau_t} - \overline{\mathbf{Z}})$ and the second term is $o_p(1)O_p(1) = o_p(1)$. By Goncalves and de Jong (2003, theorem 2), it follows that the asymptotic distribution of

$$n^{-1/2}\sum_{t=1}^{n}(\mathbf{Z}_{\tau_t} - \overline{\mathbf{Z}}) \quad \text{equals that of} \quad n^{-1/2}\sum_{t=1}^{n}\mathbf{Z}_t,$$

which is given by $N_{K_m}(0, \boldsymbol{\Omega})$ according to Lemma 4.

Consider the Taylor expansion of r, $r(\widehat{\boldsymbol{\beta}}^*) = r(\widehat{\boldsymbol{\beta}}) + \dot{r}(\widetilde{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}}^* - \widehat{\boldsymbol{\beta}})$, where $\widetilde{\boldsymbol{\beta}}$ lies between $\widehat{\boldsymbol{\beta}}^*$ and $\widehat{\boldsymbol{\beta}}$, then by the continuity of $\dot{r}$ and the fact that $\widehat{\boldsymbol{\beta}} \xrightarrow{a.s.} \boldsymbol{\beta}_0$ we have that $\dot{r}(\widetilde{\boldsymbol{\beta}}) \xrightarrow{a.s.} \dot{r}(\boldsymbol{\beta}_0)$, and hence the asymptotic distribution of $n^{1/2}[r(\widehat{\boldsymbol{\beta}}^*) - r(\widehat{\boldsymbol{\beta}})]$ equals that of $\dot{r}(\boldsymbol{\beta}_0)n^{1/2}(\widehat{\boldsymbol{\beta}}^* - \widehat{\boldsymbol{\beta}})$, which is given by $\tilde{Z} = \dot{r}(\boldsymbol{\beta}_0)Z$, where $Z \sim N_{K_m}(0, \boldsymbol{\Omega})$. By the continuous mapping theorem we find that

$$g(n^{1/2}[r(\widehat{\boldsymbol{\beta}}^*) - r(\widehat{\boldsymbol{\beta}})], \widehat{\boldsymbol{\psi}}^*) \xrightarrow{d} g(\tilde{Z}, \boldsymbol{\psi}_0),$$

since g is assumed to be continuous almost everywhere on the support of the limit distribution. ■

References

BARBIERI, M. M., AND J. O. BERGER (2002): "Optimal Predictive Model Selection," *Mimeo, Duke University*.

CAMPBELL, J., AND R. SHILLER (1988): "Stock Prices, Earnings and Expected Dividents," *Journal of Finance*, 46, 661–676.

DANILOV, D. L., AND J. R. MAGNUS (2001): "On the Harm That Pretesting Does," *CentER working paper, Tilburg University*.

DIEBOLD, F. X., AND R. S. MARIANO (1995): "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics*, 13, 253–263.

DOPPELHOFER, G., R. MILLER, AND X. X. SALA-I-MARTIN (2000): "Determinants of Long-Term Growth: A Bayesian Averaging of Classical Estimates (BACE) Approach," *NBER Working Paper 7750*.

FAMA, E., AND K. FRENCH (1989): "Business Conditions and Expected Returns on Stocks and Bonds," *Journal of Financial Economics*, 25, 23–49.

FOSTER, F. D., T. SMITH, AND R. E. WHALEY (1997): "Assessing Goodness-of-Fit of Asset Pricing Models: The Distribution of the Maximal R^2," *Journal of Finance*, 52, 591–607.

GONCALVES, S., AND R. DE JONG (2003): "Consistency of the Stationary Bootstrap under Weak Moment Conditions," *Economics Letters, forthcoming*.

GONCALVES, S., AND H. WHITE (2003): "Maximum Likelihood and the Bootstrap for Non-linear Dynamic Models," *Journal of Econometrics, Forthcoming*.

HANSEN, P. R. (2001): "A Test for Superior Predictive Ability," Brown University, Economics Working Paper. 2001-06
`http://www.econ.brown.edu/fac/Peter_Hansen`.

——— (2003): "Asymptotic Tests of Composite Hypotheses," Brown University, Department of Economics Working Paper, 03-09.
`http://www.econ.brown.edu/fac/Peter_Hansen`.

HODRICK, R. J. (1992): "Divident Yields and Expected Stock Returns: Alternative Procedures for Inference and Measurement," *Review of Financial Studies*, 5, 357–386.

INOUE, A., AND L. KILIAN (2002): "Bootstrapping Autoregressive Processes with Possible Unit Roots," *Econometrica*, 70, 377–391.

KIRBY, C. (1997): "Measuring the Predictability Variation in Stock and Bond Returns," *Review of Financial Studies*, 10, 579–630.

KUNSCH, H. R. (1989): "The Jackknife and the Bootstrap for General Stationary Observations," *Annals of Statistics*, 17, 1217–1241.

LEAMER, E. E. (1983): "Let's Take the con Out of Econometrics," *American Economic Review*, 73, 31–43.

——— (1985): "Sensitivity Analyses Would Help," *American Economic Review*, 75, 308–313.

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

LEVINE, R., AND D. RENELT (1992): "A Sensitivity Analysis of Cross-Country Growth Regressions," *American Economic Review*, 82, 942–963.

LO, A. W., AND A. C. MACKINLAY (1990): "Data-Snooping Biases in Test of Financial Asset Pricing Models," *The Review of Financial Studies*, 3, 431–467.

MERTON, R. (1987): "On the Current State of The Stock Market Rationality Hypothesis," in *Macroeconomics and Finance: Essays in Honor of Franco Modigliani*, ed. by R. Dornbusch, S. Fisher, and J. Bossons. M.I.T. Press Cambridge.

PATELIS, A. D. (1997): "Stock Return Predictability and The Role of Monetary Policy," *Journal of Finance*, 52, 1951–1972.

PESARAN, M. H., AND A. TIMMERMANN (1995): "Predictability of Stock Returns: Robustness and Economic Significance," *Journal of Finance*, 50, 1201–1228.

POLITIS, D. N., AND J. P. ROMANO (1994): "The Stationary Bootstrap," *Journal of the American Statistical Association*, 89, 1303–1313.

RENCHER, A., AND F. PUN (1980): "Inflation of R^2 in Best Subset Regressions," *Technometrics*, 22, 49–53.

ROSS, S. A. (1989): "Regression to the Max," Working paper, Yale University.

SALA-I-MARTIN, X. X. (1997a): "I Just Ran Four Million Regressions," *NBER Working Paper 6252*.

——— (1997b): "I Just Ran Two Million Regressions," *American Economic Review*, 87, 178–183.

VUONG, Q. H. (1989): "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses," *Econometrica*, 57, 307–333.

WEST, K. D. (1996): "Asymptotic Inference About Predictive Ability," *Econometrica*, 64, 1067–1084.

WHITE, H. (2000): "A Reality Check for Data Snooping," *Econometrica*, 68, 1097–1126.

Table 1: List of Candidate Regressors

Variable Name	Description	Source	Lags
ExRet	Excess return (Monthly Return on DJIA over the 1 month T-bill rate).	–	1,2
NumUps1	Indicator: One if previous ER>0	–	0
NumUps2	Indicator: # of positive ER of last two	–	0
NumUps5	Indicator: # of positive ER of last five	–	0
I1	1 month T-bill rate	AT	1
DI1	Change in 1 month T-bill rate	AT	1,2
I3	3 month T-bill rate	DRI	1
DI3	Change in 3 month T-bill rate	DRI	1,2
I6	6 month T-bill rate	DRI	1
DI6	Change in 6 month T-bill rate	DRI	1,2
I120	10 years Treasury composite (long term)	DRI	1
DI120	Change in 10 years Treasury composite (long term)	DRI	1,2
DMaDo	Change in mark/$	FRED	1,2
DYeDo	Change in Yen/$	FRED	1,2
Dppiic	Change in Producer Price Index-Industrial commodities	FRED	2,3
rppien	Relative Producer Price Index-Fuels and related products and power	FRED	2
dppien	Change in Producer Price Index-Fuels and related products and power	FRED	2,3
roilp	Relative Oil Price: Domestic West TX. Intermediate [Prior'82=Posted Price]	FRED	1
Doilp	Percentage change in relative Oil Price:...	FRED	1,2
DPZUC	Percentage change in: Consumer Price Index, Commodities	FRED	1,2
rcpile	Relative CPI-U: All Items Less Energy	FRED	2
DCPILE	Change in CPI-U: All Items Less Energy	FRED	2,3
RCPIEN	Relative CPI-U: Energy	FRED	2
DCPIEN	Change in CPI-U: Energy	FRED	2,3
DM1	Percentage change in M1	FRED	2
DM2	Percentage change in M2	FRED	2
DM3	Percentage change in M3	FRED	2
FFR	Fed Funds Rate	FRED	1,2
DTR	Percentage change in Total Reserves adjusted for changes in reserve requirements	FRED	2
DPZUC	Percentage change in CPI, Commodities	DRI	2,3
RPZUCD	Relative CPI: Durables	DRI	2
RFDB	Real Federal Debt: Total amount outstanding	DRI	2
DRFDB	Percentage change in Real Federal Debt: Total amount outstanding	DRI	2
DivBUS	Percentage change in Inventories: Business Suppliers	DRI	3,4
DivCAP	Percentage change in Inventories: Capital Goods Industy	DRI	3,4
MDO	Percentage change in new orders: Durable goods industi	DRI	3,4
MNO	Percentage change in new orders: Non durable goods.	DRI	3,4
DSALER	Percentage change in retail sale (seasonally adjusted)	FRED	2,3
IvSaRa	Inventory/sales ratio	FRED	3,4
IPXMCA	Capacity util. rate: Manufacturing	DRI	3,4
DLEAD	Composite index of 11 leading indicators	DRI	3,4
DLAGG	Composite index of 7 lagging indicators	DRI	3,4
DCOINC	Composite index of 4 coincident indicators	DRI	3,4
UNRATE	Unemployment Rate	FRED	2
DUNRAT	Change in unemployment rate	FRED	2,3
DUEM15	Percentage change in # of unempl. for 15wks or more	FRED	2,3
RUEM15	Percentage of population unemployed for 15wks or more	FRED	2,3
DMANEM	Percentage change in Manufacturing Employees	FRED	2,3
DUSSER	Percentage change in: Services Employment	FRED	2,3
DUSGOV	Percentage change in: Government Employment	FRED	2,3
DUSGOO	Percentage change in: Goods-Producing Industries Employment	FRED	2,3
DUSFIR	Percentage change in: Finance,Insurance & Real Estate Employment	FRED	2,3
DUSCON	Percentage change in: Construction Employment	FRED	2,3
HHSNTN	U.of Mich. Index of consumer expectation	DRI	2
DHHSNT	Change in U.of Mich. Index of consumer expectation (%-points)	DRI	2,3
PSAVE	Personal Savings Rate	FRED	2
DPSAVE	Change in Personal Savings Rate (%-points)	FRED	2,3
DPI	Percentage change in: Personal Income	FRED	2,3
DPCE	Percentage change in: Personal Consumption Expenditures	FRED	2,3
DPCES	Percentage change in: Personal Consumption Expenditures: Services	FRED	2,3
DPCEND	Percentage change in: Personal Consumption Expenditures: Nondurable Goods	FRED	2,3
DPCEDG	Percentage change in: Personal Consumption Expenditures: Durable Goods	FRED	2,3
DSPI	Percentage change in: Disposable Personal Income	FRED	2,3

The table lists the set of regressors. From left: the abbreviation, full description, data source, and the lags included. The total number of variables is 103. FRED refers to Federal Reserve Economic Data, CRSP refers to Center for Research in Security Prices, and AT to Allan Timmermann, who kindly provided me with the one month T-bill.

Table 2: Maximal R^2 and Critical Values

		5% Critical Values			
	$R^2_{\max}$	Miner (Beta)	Miner (Bootstrap)	BF (Beta)	Bootstrap
	DEPENDENT VARIABLE: MONTHLY RETURNS (DJIA)				
$k=1$	**0.0726**	0.0083	0.0127	0.0261	0.0316
$k=2$	**0.0990**	0.0129	0.0169	0.0635	0.0536
$k=3$	**0.1101**	0.0169	0.0220	0.0698	0.0618
	DEPENDENT VARIABLE: MONTHLY EXCESS RETURNS (DJIA)				
$k=1$	**0.0717**	0.0083	0.0126	0.0261	0.0321
$k=2$	**0.0957**	0.0129	0.0162	0.0635	0.0512
$k=3$	**0.1124**	0.0169	0.0181	0.0698	0.0644

The table reports the largest R^2 and 5% critical values. The upper half contains the results from the analysis of monthly returns on DJIA , and the lower half contains the results from the analysis of excess returns (returns minus the risk-free rate). Each row lists the number of regressors, the largest R^2 found in these models, and four 5% critical values. The four critical values: *Miner (Beta)* is the "single-model" critical value using the $B(\frac{k}{2}, \frac{n-k-1}{2})$ distribution; *Miner (Bootstrap)* is the "single-model" critical value estimated with the stationary bootstrap, in the model that had the largest R^2; *BF (Beta)* is the Bonferroni bound critical value, based on the $B(\frac{k}{2}, \frac{n-k-1}{2})$ distribution; and *Bootstrap* is the critical value estimated by the bootstrap implementation that take the inter-dependence of models into account.

Table 3: Best Set of Regressors

	$R^2_{\max}$	Best Set of Regressors			
		DEPENDENT VARIABLE: MONTHLY RETURNS (DJIA)			
$k=1$	**0.0726**	Constant	DI120_1	–	–
		0.6457	−4.1410	–	–
$k=2$	**0.0990**	Constant	DI120_1	DMANEM_2	–
		0.6796	−3.9935	−1.3304	–
$k=3$	**0.1101**	Constant	DI3_1	DI6_1	DMANEM_2
		0.6668	4.9331	−7.0297	−1.2962
		DEPENDENT VARIABLE: MONTHLY EXCESS RETURNS (DJIA)			
$k=1$	**0.0717**	Constant	DI120_1	–	–
		0.1676	−4.1288	–	–
$k=2$	**0.0957**	Constant	DI120_1	DMANEM_2	–
		0.2001	−3.9876	−1.2734	–
$k=3$	**0.1124**	Constant	DI120_1	FFR_1	DMANEM_2
		1.2681	−3.9576	−0.1605	−1.4230

The table reports regressors that led to the largest R^2 for models with $k = 1$, 2, and 3 regressors. The estimated coefficients are given below the names of the regressors.

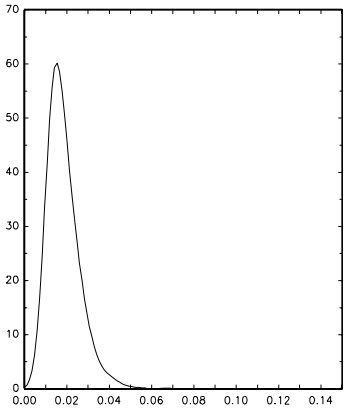

Figure 1: A kernel estimate of the bootstrap distribution of $R^2_{\max}$, over the models with two regressors and excess return as the dependent variable.

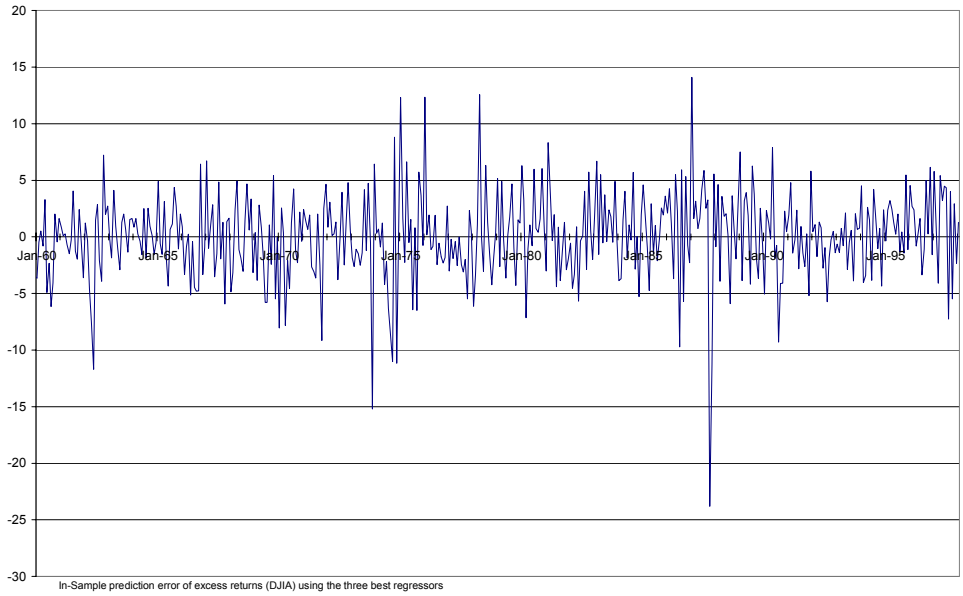

Figure 2: Prediction errors of the model with the maximal R^2.