DATA-DRIVEN MODEL EVALUATION: A TEST FOR REVEALED PERFORMANCE

JEFFREY S. RACINE AND CHRISTOPHER F. PARMETER

Abstract. When comparing two competing approximate models, the one having smallest 'expected true error' is closest to the data generating process (according to the specified loss function) and is therefore to be preferred. In this paper we consider a data-driven method of testing whether two competing approximate models, for instance a parametric and a nonparametric model, are equivalent in terms of their expected true error (i.e., their expected performance on unseen data drawn from the same data generating process). The proposed test is quite flexible with regards to the types of models and data types that can be compared (i.e., time-series, cross section, panel etc.). Moreover, by applying our method to time-series models we can overcome two of the drawbacks associated with existing approaches, namely, the reliance on only one split of the data and the need to have a sufficiently large hold-out sample in order for the test to have power. Some useful graphical summaries are also presented. Finite-sample performance and several illustrative applications are considered.

1. Introduction

Having estimated a parametric model in the course of applied data analysis, one ought naturally test for model adequacy (i.e., for correct specification). When the parametric model is rejected by the data, practitioners often turn to more flexible methods, for example, nonparametric models. But there is no guarantee that the nonparametric model one has adopted will perform any better than the parametric model that has been deemed inadequate, even though the nonparametric model may indeed exhibit an apparent marked improvement in (within-sample) fit according to a variety of metrics.[1]

This is widely appreciated in the time-series literature where out-of-sample predictive performance is an overriding concern.[2] By way of example, Medeiros, Teräsvirta & Rech (2006) consider

Date: January 12, 2009: Preliminary and Incomplete – Not to be quoted without the author's permission.
Key words and phrases. Approximate, misspecified, model selection, predictive accuracy, Data mining.
We would like to thank (but not implicate) Shahram Amini, Richard Ashley, Robert Hyndman, Nicolai Kuminoff, Esfandiar Maasoumi, Andrew Patton, Dimitris Politis and Zhiyuan Zheng for their thoughtful comments and suggestions. Racine would like to gratefully acknowledge support from Natural Sciences and Engineering Research Council of Canada (NSERC:www.nserc.ca), the Social Sciences and Humanities Research Council of Canada (SSHRC:www.sshrc.ca), and the Shared Hierarchical Academic Research Computing Network (SHARCNET:www.sharcnet.ca).

[1]Alternatively, as White (2000, page 1097) discusses, resorting to extensive specification searches runs the risk that the observed good performance of a model is not the result of superior fit but rather luck and he labels such practices 'data snooping'.

[2]Corradi & Swanson (2002, page 356) underscore the importance of this issue when they discuss "... whether simple linear models (e.g. ARIMA models) provide forecasts which are (at least) as accurate as more sophisticated nonlinear models. If this were shown to be the case, there would be no point in using nonlinear models for out-of-sample

using autoregressive neural network models (AR-NN) to model financial time-series. However, having rejected linearity, fitted an AR-NN model, and conducted a rigorous post-mortem analysis of each model's ability to predict stock returns, Medeiros et al. (2006, page 69) conclude that the "NN modelling strategy [...] is not any better than a linear model with a constant composition of variables. A nonlinear model cannot therefore be expected to do better than a linear one."

Indeed, there is no guarantee that a parametric model that passes a test for model adequacy will perform better than a nonparametric model as it is known that overspecified parametric models may perform worse than alternative specifications, including nonparametric ones. However, focusing instead on out-of-sample predictive ability may provide the applied researcher with a potential avenue for discriminating among such models. Though a literature that advocates in-sample predictive evaluation in time-series settings has recently emerged (see Inoue & Kilian 2004), this runs against the tide of a large body of literature that convincingly argues for the use of sample-splitting mechanisms whereby one randomly divides the full sample into two sub-samples, then uses one sub-sample for estimation and the other to guide predictive evaluation; see Corradi & Swanson (2007) and the references therein.

Out-of-sample predictive performance has become the metric of choice for time-series researchers; see Diebold & Mariano (1995), West (1996), West & McCracken (1998) and McCracken (2000), among others. However, to the best of our knowledge, the insights underlying this literature have as yet to permeate cross-section applications. Furthermore, as will be demonstrated, there remains scope for improving these methods. In this paper we demonstrate how the 'forecast ability' of cross-sectional models can assist practitioners who wish to discriminate one approximate model from another based upon their respective expected performance on independent data drawn from the same underlying data generating process (DGP). Thus, this paper might be viewed as the cross-sectional counterpart to the literature on forecasting and predictive ability that exists in the time-series domain, though the method we propose is somewhat more involved than these time-series approaches as will be seen. Furthermore, by using an appropriate resampling mechanism it will be seen that the proposed approach can provide an appealing alternative to popular time-series tests for predictive accuracy by overcoming what we regard as limitations associated with such tests.

In this paper we take the view that fitted statistical models are approximations,[3] a perspective that differs from that of consistent model selection which posits a finite-dimensional 'true model'. That is, in this paper we are not interested in tests that hypothesize one model being the 'true model'. Rather, our goal is instead to test whether one approximate model's expected performance is better than another on data drawn from the same DGP according to a pre-specified loss function

prediction, even if the linear models could be shown to be incorrectly specified, say based on the application of *in-sample* nonlinearity tests..." (our italics).

[3]See Hansen (2005, pgs.62-63) for an eloquent discussion of this issue.

such as square or absolute error loss. The loss function is provided by the user hence the method suggested herein is quite general.[4]

Our approach is firmly embedded in the statistics literature dealing with apparent versus true error estimation; for a detailed overview of 'apparent', 'true', and 'excess' error, we direct the reader to Efron (1982, Chapter 7). In effect, within-sample measures of fit gauge 'apparent error' which will be more optimistic than 'true error', sometimes strikingly so, since a model is selected to *fit* the data best. For a given loss function, $\ell(u)$, one might compute the expected loss, $n^{-1}\sum_{i=1}^{n}\ell(u_i)$, which provides an estimate of the apparent error arising from the modelling process. But all such within-sample measures are fallible which is why they cannot be recommended as guides for model selection; for example, R^2 does not take into account model complexity, adjusted R^2 measures are not defined for many nonparametric methods, whereas information-based measures such as AIC can be biased if the sequence of competing (parametric) models is nonnested; see Ye (1998) and Shen & Ye (2002).

The approach we advocate involves constructing the distribution function of a model's *true error* and *testing* whether the *expected* true error is statistically smaller for one model than another. This will be accomplished by leveraging *repeated* splits of the data rather than just one as is commonly done and by computing the estimated loss for the *hold-out* data for each split. At the end of the day we will conclude that one model has statistically smaller estimated expected true error than another and therefore is expected to be closer to the true DGP hence is preferred though both models are, at best, approximations.

The basic idea is, of course, not new and involves splitting the data into two independent samples of size n_1 and n_2, fitting models on the first n_1 observations, then evaluating the models on the remaining $n_2 = n - n_1$ observations using, say, average square prediction error (ASPE) (we know the y values for the evaluation data, hence this delivers an estimate of true error).[5] However, one might mistakenly favor one model when the estimate of true error is lower but this in fact simply reflects a particular division of the data into two independent subsets which may not be representative of the DGP, i.e., this can be overly influenced by which data points end up in each of the two samples. To overcome this limitation, one might consider repeating this process a large number of times, say $S = 10,000$ times, each time refitting the models on the 'training' data (the n_1 observations) and evaluating on the independent 'evaluation' data (the $n_2 = n - n_1$ hold-out observations). This repeated sample-splitting experiment will thereby produce two vectors of length S which represent

[4]This allows us to address how much more accurate one method is compared to another *on average* with respect to the chosen loss function. Indeed, this is in direct agreement with Goodwin (2007, page 334): "[...] when comparing the accuracy of forecasting methods [...] The interesting questions are, how much more accurate is method A than method B, and is this difference of practical significance?". Our approach will allow a researcher to tackle both of these questions in a simple and easily implementable framework, though we take a broader view by considering 'out-of-sample prediction' in cross-section settings and 'forecasting' in time-series ones.

[5]Readers familiar with Diebold & Mariano's (1995) test for predictive accuracy will immediately recognize this strategy.

draws from the distribution of actual ASPEs for each model.[6] These two vectors of draws can then be used to discriminate between the two models.[7] For what follows we consider a simple test of differences in means for the two distributions, but also consider simple graphical tools that will help reveal stochastic dominance relationships, if present. Given that the test is a test of whether the data at hand reveals that the predictive performance of one econometric model is statistically different from that of another, we dub the test the 'RP' test to denote 'revealed performance'.

The statistics literature on cross-validated estimation of excess error is a well-studied field ('expected excess error' is the expected amount by which the true error exceeds the apparent error). However, this literature deals with model specification within a class of models (i.e., which predictor variables should be used, whether or not to conduct logarithmic transformations on the dependent variable and so forth) and proceeds by minimizing excess error. Our purpose here is substantively different, and is perhaps most closely related to the literature on non-nested model testing (see Davidson & MacKinnon 2002). Unlike this literature, however, we are asking an inherently different question that is not the subject of interest in the non-nested literature, namely, whether the expected true error associated with one model differs *significantly* from that for another model, whether nested or not.

Our test is quite flexible with regards to the types of models that can be compared. The flexibility of the test stems from the fact that it does not require both models to be of the same type (e.g., of parametric versus parametric type) nor is it limited by the type of data at hand (i.e., time-series, cross section, panel etc.). In fact, while our focus here is on regression models, the insight here can be extended to predictions from count data or limited dependent variable models, probability models, quantile frameworks and so forth, i.e., any model for which we have a response and set of explanatory variables.[8] Moreover, by applying our method to time-series models we can overcome two of the drawbacks associated with existing approaches, namely, the reliance on a single split of the data and the need to have a sufficiently large hold-out sample in order for the test to have power.

The rest of this paper proceeds as follows. Section 2 outlines our basic approach and defines the framework for our proposed test. Section 3 conducts several simulation exercises to assess the performance of the proposed approach when the DGP is known. Section 4 presents several empirical examples. Section 5 presents some concluding remarks.

[6]Clearly, for (strictly) stationary dependent processes we cannot use sample splitting directly, however, we can use resampling methods that are appropriate for such processes. Used properly, each resample will respect dependence in the original series and can itself be split; see Politis & Romano (1992) by way of example.

[7]For instance, we might perform a test of the hypothesis that the mean ASPE ('expected true error') for the $S = 10,000$ splits is equal (less than or) for two models against a one-sided alternative (greater than) in order to maximize power. Or, one could test for stochastic dominance of one distribution over the other.

[8]Additionally, while we do not discuss it further, our test is not restricted to the case where the dependent variable is identical across models. One could use the insight of Wooldridge (1994) to transform the predictions from one model to that of another where the dependent variable was transformed (monotonically).

2. Methodology

The method we describe here is closest in spirit to the original application of cross-validation in which the data set is randomly divided into two halves, the first of which is used for model fitting and the second for cross-validation where the regression model fitted to the first half of the data is used to predict the second half. The more common modern variant in which one leaves out one data point at a time, fits the model to the remaining points, then takes the average of the prediction errors (each point being left out once) yielding a cross-validated measure of true error, has been widely studied, and we direct the interested reader to Stone (1974), Geisser (1975), and Wahba & Wold (1975) for detailed descriptions of this method. It is noteworthy that White (2000, page 1108) argues that cross-validation represents a "more sophisticated use of 'hold-out' data" and indicates that this "is a fruitful area for further research". Our approach indeed supports this claim as we demonstrate that cross-validation can lead to dramatic power improvements over existing, single-split techniques commonly used in the applied times-series literature and elsewhere.

Though we shall begin with *iid* data and pure sample splitting (i.e., resampling *without* replacement), we shall see how that the same intuition carries over to a variety of dependent data structures as well but, as alluded to above, for (strictly) stationary dependent processes we cannot use sample splitting directly, though we can use resampling methods that are appropriate for such processes. With some care, each resample will respect dependence in the original series and can itself be split (e.g., Politis & Romano 1992) thereby allowing us to apply our method in time-series settings.

In our regression problem the data consists of pairs $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$ where X_i is a $1 \times p$ vector of predictor variables and Y_i is a real-valued response variable. For our purposes we presume that $Z_i = (X_i, Y_i)$ represent random draws from a (strictly) stationary ergodic process with unknown distribution function F defined on $\mathcal{H} = \mathbb{R}^{p+1}$,

$$Z_1, Z_2, \ldots, Z_{n_1} \sim F. \tag{1}$$

We observe $Z_1 = z_1, Z_2 = z_2, \cdots, Z_{n_1} = z_{n_1}$ and for what follows we let $Z^{n_1} = (Z_1, Z_2, \ldots, Z_{n_1})$ and $z^{n_1} = (z_1, z_2, \ldots, z_{n_1})$. Having observed $Z^{n_1} = z^{n_1}$ we fit a regression model which will be used to predict some 'future' values of the response variable, which we denote

$$\hat{g}_{z^{n_1}}(x^{n_2}), \tag{2}$$

where the superscript n_2 indicates a new set of observations, $z^{n_2} = (z_{n_1+1}, z_{n_1+2}, \ldots, z_n)$, which are distinct from $z^{n_1} = (z_1, z_2, \ldots, z_{n_1})$ where $n_2 = n - n_1$. By way of example, simple linear regression would provide $\hat{g}_{z^{n_1}}(x^{n_2}) = x^{n_2}\hat{\beta}_{n_1}$ where $\hat{\beta}_{n_1} = (x^{n_1 T} x^{n_1})^{-1} x^{n_1 T} y^{n_1}$, T denotes transpose, and $y^{n_1} = (y_1, y_2, \ldots, y_{n_1})$.

We are interested in estimating a quantity known as 'expected true error' (Efron 1982, page 51).[9] Following Efron (1982), we first define the 'true error' to be

$$E_{n_2,F}\left[\ell\left(Y^{n_2}-\hat{g}_{Z^{n_1}}(X^{n_2})\right)\right]. \tag{3}$$

The notation $E_{n_2,F}$ indicates expectation over the new point(s)

$$Z_{n_1+1}, Z_{n_1+2}, \ldots, Z_n \sim F, \tag{4}$$

independent of $Z_1, Z_2, \cdots, Z_{n_1}$, the variables which determine $\hat{g}_{n_2}(\cdot)$ (we refer to Z^{n_1} as the 'training set', terminology borrowed from the literature on statistical discriminant analysis). Next, we define 'expected true error,'

$$E\left(E_{n_2,F}\left[\ell(\cdot)\right]\right), \tag{5}$$

the expectation over all potential regression surfaces $\hat{g}_{z^{n_2}}(\cdot)$, for the selected loss function. When comparing two approximate models, the model possessing lower 'expected true error' will lie closest to the true DGP hence will deliver a better approximation and would therefore be preferred in applied settings.

A realization of 'true error' (3) based upon the observed $z^{n_2} = (z_{n_1+1}, z_{n_1+2}, \ldots, z_n)$, is given by

$$\frac{1}{n_2}\sum_{i=n_1+1}^{n_2}\ell\left(y_i-\hat{g}_{z^{n_1}}(x_i^{n_2})\right), \tag{6}$$

an average prediction error which, for square error loss, we denote ASPE ('average square prediction error'). If we consider S such splits of the data, we can then construct the empirical distribution function (EDF) of loss.

Given two competing models and each model's respective loss EDF, and we can then use the respective EDFs to determine whether one model has statistically significantly lower expected true error than another. Note that here we have transformed the problem into a two sample problem where we wish to test for equivalence of expected true error defined in (5) based upon two vectors of realizations of true errors defined in (6).

2.1. **The Empirical Distribution of True Error.** Suppose we arbitrarily denote one approximate model 'Model A' and the other 'Model B'. For the sake of concreteness, let us presume one was interested in comparing, say, a nonparametric kernel regression model ('Model A') to a parametric regression model ('Model B'). We first consider the case where the data represent independent draws from the underlying DGP. When the data represent independent draws we could proceed as follows:

(i) Resample without replacement pairwise from $z = \{x_i, y_i\}_{i=1}^n$ and call these $z_* = \{x_i^*, y_i^*\}_{i=1}^n$.
(ii) Let the first n_1 of the resampled observations form a training sample, $z_*^{n_1} = \{x_i^*, y_i^*\}_{i=1}^{n_1}$ and the remaining $n_2 = n - n_1$ observations form an evaluation sample, i.e., $z_*^{n_2} = \{x_i^*, y_i^*\}_{i=n_1+1}^n$.

[9] Efron (1982, page 51) considers estimation of 'expected excess error', while we instead consider estimation of 'expected true error'.

(iii) Holding the degree of smoothing[10] (i.e., the bandwidth vector scaling factors) of Model A and the functional form of Model B fixed (i.e., at that for the full sample), fit each model on the training observations ($z_*^{n_1}$) and then obtain predicted values from the evaluation observations ($z_*^{n_2}$) that were not used to fit the model.

(iv) Compute the ASPE of each model which we denote $\text{ASPE}^A = n_2^{-1}\sum_{i=n_1+1}^{n}(y_i^* - \hat{g}^A_{z^{n_2}}(x_i^*))^2$ and $\text{ASPE}^B = n_2^{-1}\sum_{i=n_1+1}^{n}(y_i^* - \hat{g}^B_{z^{n_2}}(x_i^*))^2$.

(v) Repeat this a large number of times, say, $S = 10,000$, yielding S draws, $\{\text{ASPE}_s^A, \text{ASPE}_s^B\}_{s=1}^S$. We refer to the respective empirical distributions as $\hat{F}_S^A$ and $\hat{F}_S^B$ where each places mass $1/S$ at ASPE_s^A and ASPE_s^B.

Step (i), which involves resampling without replacement from $z = \{x_i, y_i\}_{i=1}^n$, is valid for heteroskedastic processes however it does presume independence among the draws. That is, by resampling (x_i, y_i) pairs we avoid resorting to, for instance, the 'wild bootstrap' which is a residual-based bootstrap that admits heteroskedastic errors. However, in a time-series context independent pairwise resampling is clearly inappropriate. For one thing, univariate time-series models are quite popular but require a different treatment as we need to respect dependence in the series itself. In a time-series context, it might appear that there is only one possible split of the data, $\{z_i\}_{i=1}^t$ and $\{z_i\}_{i=t_1}^n$, this split underlies many tests for predictive accuracy (or forecast equality) such as Diebold & Mariano's (1995) test. But, there is nothing to preclude conducting step (i) (resampling) with time-series data; we just need to use an appropriate resampling methodology.

In the context of time-series prediction ('forecasting'), resampling methods are widely used. For instance, Corradi & Swanson (2002) propose a consistent test for nonlinear predictive accuracy for nested models where interest lies in testing whether the null model can outperform the nesting alternative model based upon "real-time forecasts" (i.e., one-step recursive forecasts for period $t+1$, $t+2$ and so on) and one split of the data. Corradi & Swanson (2004) examine finite-sample properties of their (2002) test where critical values are based on application of the block bootstrap. Corradi & Swanson (2004, Tables 1 and 2) employ manually set fixed block lengths and they note that the value of the test statistic(s) under consideration and the resulting power properties vary dramatically as the block length is changed. There is no reason to require the user to set

[10] A 'scaling factor' refers to the unknown constant c in the formula for the optimal bandwidth, $h_{opt} = cn^{-1/(4+p)}$. Cross-validation can be thought of as a method for estimating the unknown constant c, where c is independent of the sample size n. This constant can then be rescaled for samples of differing size drawn from the same DGP thereby ensuring that the same degree of smoothing is applied to the full sample and the subsample (see Racine 1993). The rationale for so doing is as follows. Think of estimating a univariate density function where the data represent independent draws from the $N(0,1)$ distribution. The optimal bandwidth in this case is known to be $h_{opt} = 1.059n^{-1/5}$. If $n = 200$ then $h_{opt} = 0.3670$ while if $n = 100$ then $h_{opt} = 0.4215$. Cross-validation delivers an estimate of h_{opt} for a sample of size n, i.e. $\hat{h} = \hat{c}n^{-1/5}$, while it can be shown that $(\hat{h} - h_{opt})/h_{opt} \to 1$ asymptotically (see Stone 1984). If you don't rescale the cross-validated bandwidth for subsamples of size $n_1 < n$ (i.e., adjust $\hat{h}$ when n falls to n_1) then you are in fact doing a different amount of smoothing on subsamples of size $n_1 < n$ (i.e., $h = 0.3670$ will undersmooth when $n_1 < 200$, so the estimate based on $h = 0.3670$ and $n_1 < 200$ will be overly variable). But, by using $\hat{c}$ corresponding to the cross-validated bandwidth for the full sample, $\hat{h}$, we can ensure that the same degree of smoothing is applied to the subsamples of size $n_1 < n$ and the full sample of size n. This keeps the baseline nonparametric model fixed at that for the full sample, in the same way that we hold the functional form of the parametric model fixed at that for the full sample.

block lengths manually, however, just as there is no reason to require users to manually specify bandwidths for kernel estimation; automatic methods having desirable statistical properties have recently been proposed.

In what follows, we shall exploit recent advances in time-series resampling methodology, and use geometric ('stationary') block bootstrapping to generate a bootstrap replication of the series of size n which then can itself be split into two samples of size n_1 and n_2 (i.e., step (ii)) thereby preserving the dependence structure present in the full sample. That is, in a time-series setting, we simply modify step (i), which involves resampling without replacement from $z = \{x_i, y_i\}_{i=1}^n$, with a time-series bootstrap based on automatic block length selection where we resample from, say, $z = \{y_i\}_{i=1}^n$. By way of illustration, we elect to use the method of Politis & Romano (1992) for which Politis & White (2004) recently proposed a fully automatic method for choosing the block length that has excellent finite-sample properties.[11] This bootstrap preserves the underlying dependence structure by resampling the data in blocks of random length, where the lengths are derived from a geometric distribution, hence the name. See both Davison & Hinkley (1997, pp.401-408) and Lahiri (2003, sections 2.7.2 and 3.3) for more on the theoretical underpinnings underlying the geometric bootstrap.[12] In a time-series setting, we can modify (i)-(v) such that, when the data represent draws from (strictly) stationary ergodic time-series process, we therefore proceed as follows:

(i) Apply the stationary bootstrap to resample from $z = \{y_i\}_{i=1}^n$ and call these $z_* = \{y_i^*\}_{i=1}^n$.

(ii) Let the first n_1 of the resampled observations form a training sample, $z_*^{n_1} = \{y_i^*\}_{i=1}^{n_1}$ and the remaining $n_2 = n - n_1$ observations form an evaluation sample, i.e., $z_*^{n_2} = \{y_i^*\}_{i=n_1+1}^n$.

(iii) Holding the degree of smoothing of Model A and the functional form of Model B fixed (i.e., at that for the full sample), fit each model on the training observations ($z_*^{n_1}$) and then generate predictions for the n_2 evaluation observations.

(iv) Compute the ASPE of each model which we denote $\text{ASPE}^A = n_2^{-1} \sum_{i=n_1+1}^n (y_i^* - \hat{g}_{z^{n_2}}^A(y_{i-1}^*, \dots))^2$ and $\text{ASPE}^B = n_2^{-1} \sum_{i=n_1+1}^n (y_i^* - \hat{g}_{z^{n_2}}^B(y_{i-1}^*, \dots))^2$.

(v) Repeat this a large number of times, say, $S = 10,000$, yielding S draws, $\{\text{ASPE}_s^A, \text{ASPE}_s^B\}_{s=1}^S$. We refer to the respective empirical distributions as $\hat{F}_S^A$ and $\hat{F}_S^B$ where each places mass $1/S$ at ASPE_s^A and ASPE_s^B.

We can now proceed to use $\hat{F}_S^A$ and $\hat{F}_S^B$ to discriminate between models. At this stage we point out that the choice $S = 1$ is typically used to discriminate among time-series models, i.e., one split only of the data is the norm. By way of example, the popular time-series test for predictive accuracy of Diebold & Mariano (1995) is based on only one split, hence attention has shifted to determining

[11] See Patton, Politis & White (2008) for a correction to several of the results in Politis & White (2004).

[12] Our choice of the stationary block bootstrap is for expositional purposes. In practice we recommend that the user employ a bootstrap appropriate for the type of dependence apparent in the data. For example, additional types of bootstraps are the Markov conditional bootstrap Horowitz (2004), the circular bootstrap Politis & Romano (1992) and the sieve bootstrap Bühlmann (1997); see Lahiri (2003) for an up-to-date and detailed coverage of available block resampling schemes. One can easily implement a variety of block bootstrap procedures by using the tsboot() command available in the boot package (Canty & Ripley 2008) in R (R Development Core Team 2008).

how large n_2 need be (e.g., see Ashley 2003), while Anglin & Gençay (1996) consider one split of their data with $n_2 = 10, 20$. One might, however, be worried about basing inference on only one split simply because one might mistakenly favor one model over another when this simply reflects a particular division of the data into two independent subsets that may not be representative of the DGP, i.e., this can be overly influenced by which data points end up in each of the two samples.

However, by instead basing inference on $S \gg 1$ (i.e., averaging over a large number of such splits), we can control for mistakes arising from divisions of the data that are not representative of the DGP. In fact, it will be seen that the power of our test increases with S, which is obvious in hindsight. Furthermore, by averaging over a large number of splits, we can base inference on much smaller evaluation sample sizes (i.e., n_2) thereby taking maximal advantage of the estimation data which would be particularly advantageous in time-series settings. Ashley (2003) clearly illustrates this dilemma in the $S = 1$ time-series context by highlighting that one may need n_2 to be quite large in order for such tests to have power; the dilemma is that for a time-series of fixed length n, increasing $n_2 = n - n_1$ means that the models are less efficiently estimated since they are based on fewer observations. Our approach will be seen to effectively overcome this limitation.

2.2. **Validity of the Bootstrap.** We now consider conditions that justify our use of the bootstrap for obtaining valid approximations to the unknown loss distributions for two competing approximate models, which we denote F^A and F^B, respectively. For what follows we leverage Lemma A.3 and Theorem 2.3 in White (2000) to establish consistency of our bootstrap approach. The conditions required (for competing parametric models) involve assumptions on the data, parameter estimates, behavior of the bootstrap, properties of the loss function, and some additional regularity conditions. Before proceeding we note that the proof we provide is for the time-series method we describe above. However, for *iid* data the automatic block length selection mechanism and geometric bootstrap that we use for our time-series approach (Patton et al. (2008)) will in fact deliver an appropriate bootstrap for independent data since it will select a block length of one in probability in these settings hence collapses to the mechanism we consider for independent data. That is, the proof will cover both cases considered above as will the implementation. For concreteness, we focus our theoretical arguments on the case where the competing models are both of parametric form (but potentially nonlinear). Extensions to semiparametric and nonparametric estimators are easily handled with (minor) modifications to the requisite assumptions listed below. As the conditions we impose involve theoretical arguments described in three separate papers, we shall outline each set of assumptions in turn and cite sources accordingly.

We begin with an assumption given in Politis & Romano (1994) that is required to demonstrate consistency of the stationary bootstrap under a range of settings. For what follows we have $\ell_s = \ell(u_s)$ where the index of s follows from the context. β^* represents an unknown parameter vector.

Assumption 2.1.

(i) Let q denote the probability of the geometric distribution used for the stationary bootstrap (q is equivalent to one over the block length). Assume that $q \to 0$ and that $nq \to \infty$ as $n \to \infty$.

(ii) Let $Z_1, Z_2, \ldots,$ be a strictly stationary process with $E|Z_1|^{6+\eta} < \infty$ for some $\eta > 0$.

(iii) Let $\{Z_n\}$ be α-mixing with $\alpha_Z(k) = O(k^{-r})$ for some $r > 3(6+\eta)/\eta$.

Assumption 2.1(i) establishes the rate at which the block length in the stationary bootstrap can grow. Assumptions 2.1(ii) and 2.1(iii) are required to ensure the data behaves in a manner consistent with the theoretical arguments of both Politis & Romano (1994) and White (2000). Of course, in cross-section settings these conditions are automatically satisfied. Note that Assumption 2.1 is the same as that used by Politis & Romano (1994) for much of their theoretical work in this area (see Theorems 2-4, Politis & Romano 1994).

Additionally, we require assumptions 1-4 in West (1996). We restate these and label them jointly as Assumption 2.2.

Assumption 2.2.

(i) Let the loss function be measurable and second order continuously differentiable at β^*. Additionally, the matrix of second order derivatives is dominated by m_n where $E[m_n] < D$ for $D < \infty$.

(ii) Let the parameter estimates be linear combinations of orthogonality conditions used to identify the response. More formally we have that (for a parametric regression model, $y_s = X_s\beta + \varepsilon_s$) $\hat{\beta} - \beta^* = B(n)H(n)$ where $B(n) \xrightarrow{a.s.} B = (EX^TX)^{-1}$ and $H(n) = n^{-1}\sum_{s=1}^{n} h_s(\beta^*) = n^{-1}\sum_{s=1}^{n} X_s\varepsilon_s$. Here $h_s(\cdot)$ is our orthogonality condition.

(iii) Let

$$\ell_s = \ell(,\beta^*), \quad \ell_{s,\beta} = \frac{\partial \ell_s}{\partial \beta}(,\beta^*), \quad F = E[\ell_{s,\beta}].$$

(a) For some $d > 1$, $\sup_s E||[vec(\ell_{s,\beta})', \ell_s', h_s']'||^{4d} < \infty$, where $||\cdot||$ signifies the Euclidean norm. (b) $[vec(\ell_{s,\beta} - F)', (\ell_s - E[\ell_s])', h_s']'$ is strong mixing with mixing coefficient of size $-3d/(d-1)$. (c) $[vec(\ell_{s,\beta})', \ell_s', h_s']'$ is covariance stationary. (d) $S_{\ell\ell}$ is positive definite where $S_{\ell\ell} = \sum_{j=-\infty}^{\infty} \Gamma_{\ell\ell}(j)$ and $\Gamma_{\ell\ell}(j) = E[(\ell_s - E[\ell_s])(\ell_{s-j} - E[\ell_s])']$.

(iv) Let $n_1, n_2 \to \infty$ as $n \to \infty$ and let $\lim_{n\to\infty}(n_2/n_1) = c$, for $0 \leq c \leq \infty$.

Assumption 2.2(i) ensures that the loss function is well behaved in a neighborhood of a specified parameter value. Essentially, the loss function evaluated at the prediction errors needs to be bounded and satisfy certain moment conditions in order to use White's (2000) bootstrap theory. As noted by West (1996), Assumption 2.2(ii) does not assume that either ε or $X\varepsilon$ is serially uncorrelated. Assumption 2.2(iii) is used to pin down the behavior of the mean of the losses for a particular model by suitable application of a law of large numbers applicable to mixing processes (see Section 3.4, White 2001). Assumption 2.2(iv) is needed to invoke asymptotic arguments related to either the estimation sample size (n_1) or the evaluation sample size (n_2).

In order to invoke either Lemma A.3 or Theorem 2.3 of White (2000) we need two additional conditions. In White (2000) these are Assumption A.5 and Assumption C. We state them here for convenience.

Assumption 2.3.

(i) Let the spectral density of $[(\ell_s - E\ell_s)', h_s' B']'$ *where* $\ell_s = \ell(y_s - \hat{y}_s)$*, at frequency zero, multiplied by a scale factor, be positive definite.*

(ii) Let the parameter estimates $(\hat{\beta})$ *obey a law of iterated logarithm.*

Assumption 2.3(ii) is required to bound a pseudo-studentized term involving $\hat{\beta}$ in White's (2000) Theorem 2.3.

These conditions are sufficient to establish that the bootstrap distribution of any (parametric) candidate model's evaluation sample loss is consistent for the distribution of expected true error, which we now state formally.

Theorem 2.1. *Under Assumptions 2.1, 2.2 and 2.3, the stationary bootstrap estimates of the distributional laws* F^A *and* F^B*, denoted* $\hat{F}^A$ *and* $\hat{F}^B$*, converge in probability to* F^A *and* F^B*.*

Proof. Given that Assumptions 2.1, 2.2 and 2.3 are identical to those in White (2000), we can invoke his Theorem 2.3 (which follows immediately from Lemma A.3) directly to achieve the result. We mention that Theorem 2.3 follows under the condition that the objective function used for estimation and loss function are equivalent. This is not a deterrent as Corradi & Swanson (2007, Proposition 1, page 77) generalize the results of White (2000) for the case where the loss function differs from the objective function used to obtain the parameter estimates. In our work, and certainly for most applications, they are identical. □

Theorem 2.1 allows us to therefore implement a variety of two-sample tests to assess revealed performance (pairwise) across a set of candidate models. In what follows we consider a simple t-test for equality in means to assess whether one distribution dominates the other (i.e., test equality (less than or equal) of means against the alternative hypothesis that the true difference in means is greater than zero). Formally, we state the null and alternative as

$$H_0 : E\left(E_{n_2,F^A}\left[\ell(\cdot)\right]\right) - E\left(E_{n_2,F^B}\left[\ell(\cdot)\right]\right) \leq 0$$

and

$$H_A : E\left(E_{n_2,F^A}\left[\ell(\cdot)\right]\right) - E\left(E_{n_2,F^B}\left[\ell(\cdot)\right]\right) > 0$$

which arises directly from our notation in (5).

This is, of course, not the only test available to practitioners. One might prefer, say, the Mann-Whitney-Wilcoxon test (i.e., test equality (less than or equal) of locations against the alternative hypothesis that the true location shift is greater than zero); see Bauer (1972). Or perhaps one might undertake a more sophisticated test for, say, first-order stochastic dominance (e.g., Davidson & Duclos 2000). We argue that this is not needed in the present context and a simple test for equality of locations and summary plots of the vectors of ASPEs is more than sufficient for our

purposes. Indeed, one of the appealing aspects of the proposed approach lies in its simplicity, though nothing would preclude the practitioner from considering additional tests in this setting as they will all be based on $\hat{F}_S^A$ and $\hat{F}_S^B$ which have been pre-computed and are consistent given Theorem 2.1.

We now proceed to some Monte Carlo simulations designed to assess the finite-sample performance of the proposed method.

3. Monte Carlo Simulations

3.1. **Finite-Sample Performance: Cross-Sectional Data.** We begin with a series of simulations that assess the finite-sample performance of the proposed data in the presence of cross-sectional data, and we consider a DGP of the form

$$y_i = 1 + x_{i1} + x_{i2} + \delta\left(x_{i1}^2 + x_{i2}^2\right) + \varepsilon_i, \tag{7}$$

where $X \sim U[-2, 2]$ and $\varepsilon \sim N(0, 1)$. By setting $\delta = 0$ we simulate data from a linear model and by setting $\delta \neq 0$ we simulate data from a quadratic model with varying strength of the quadratic component.

For what follows, we estimate a range of parametric models starting with one that is linear in X_1 and X_2 then ones that include higher order polynomials in X_1 and X_2 along with local constant and local linear nonparametric models. We consider testing whether the local linear nonparametric specification is preferred to a local constant nonparametric specification and each of the parametric specifications that are linear in X_1 and X_2 ($P = 1$), linear and quadratic in X_1 and X_2 ($P = 2$) and so forth through models that have quintic specifications. Clearly our test is designed to compare two models only, hence we intend this exercise to be illustrative in nature. Models with $P > 2$ are therefore *overspecified* parametric models. The null is that a particular model has true error (as measured by ASPE) that is lower than or equal to the local linear (LL) model, the alternative that it has ASPE that exceeds that for the LL model. The nonparametric models use cross-validated bandwidth selection while the parametric models are fit by the method of least squares.

For what follows we set $n = 200$, $S = 1,000$ or $10,000$ (to investigate the impact of increasing the number of sample splits), consider a range of values for n_2, and report empirical rejection frequencies at the 5% level in Table 1. High rejection frequencies indicate that the LL model is preferred, low that it is not. That is, a high rejection frequency (e.g., the local constant (LC) with $\delta = 0$, 0.960, $n_2 = 25$) indicates that the LC model is deemed to be inferior to the LL model on ASPE grounds when the DGP is in fact linear (only 4% of the time in this set of simulations is the LC estimator preferred to the LL estimator). A small rejection frequency (e.g., the parametric linear model ($P = 1$) with $\delta = 0$, 0.106, $n_2 = 25$) indicates that a large percentage of times the linear model is deemed to be superior to the LL model (89.4% of the time in this set of simulations) on ASPE grounds when the DGP is in fact linear.

Table 1 reveals a number of interesting features. For example, overspecified parametric models (that would be expected to pass tests for correct parametric specification) can be dominated by

TABLE 1. Each entry represents rejection frequencies 5% level for the test that the LL model has predictive accuracy equal to that for each model in the respective column heading, rejecting when the model in the respective column heading has improved predictive accuracy.

$n = 200$, $n_2 = 5$, $S = 1,000$.

δ	LC	P=1	P=2	P=3	P=4	P=5
0.0	0.749	0.023	0.036	0.077	0.336	0.611
0.2	0.422	0.761	0.015	0.023	0.048	0.105
0.4	0.294	1.000	0.003	0.006	0.012	0.017

$n = 200$, $n_2 = 25$, $S = 1,000$.

δ	LC	P=1	P=2	P=3	P=4	P=5
0.0	0.960	0.106	0.377	0.769	0.899	0.937
0.2	0.836	0.929	0.049	0.111	0.222	0.457
0.4	0.739	1.000	0.011	0.020	0.038	0.086

$n = 200$, $n_2 = 50$, $S = 1,000$.

δ	LC	P=1	P=2	P=3	P=4	P=5
0.0	0.971	0.125	0.694	0.891	0.948	0.969
0.2	0.931	0.961	0.061	0.143	0.362	0.637
0.4	0.824	0.999	0.008	0.018	0.049	0.112

$n = 200$, $n_2 = 100$, $S = 1,000$.

δ	LC	P=1	P=2	P=3	P=4	P=5
0.0	0.986	0.105	0.854	0.951	0.974	0.984
0.2	0.933	0.921	0.035	0.213	0.584	0.853
0.4	0.866	0.997	0.000	0.004	0.027	0.161

$n = 200$, $n_2 = 150$, $S = 1,000$.

δ	LC	P=1	P=2	P=3	P=4	P=5
0.0	0.959	0.045	0.852	0.920	0.952	0.971
0.2	0.888	0.632	0.066	0.489	0.839	0.941
0.4	0.746	0.933	0.000	0.001	0.102	0.681

the nonparametric local linear specification (e.g., $P = 4$ and $P = 5$, $\delta = 0.0, 0.1$), which may be surprising to some. Furthermore, the power of the proposed approach to discriminate against incorrectly specified parametric models approaches one as δ increases (the column with heading $P = 1$/LM) suggesting that the test can correctly reveal that a nonparametric model is preferred to an incorrectly underspecified parametric model. Also, the results of the test appear to stabilize after $n_2 = 25$, indicating that the size of the hold-out sample is not a crucial parameter to be set by the user; see Ashley (2003) for more on the appropriate size of the hold-out sample for forecasting in time-series domains. It is easy enough for the user to investigate the stability of their results with respect to n_2, and we encourage such sensitivity checks in applied settings.

Comparing the corresponding entries in Table 1 ($S = 1,000$) to Table 2 ($S = 10,000$), we observe that power increases with S as expected. This suggests that when one fails to reject the null it

TABLE 2. Each entry represents rejection frequencies 5% level for the test that the LL model has predictive accuracy equal to that for each model in the respective column heading, rejecting when the model in the respective column heading has improved predictive accuracy.

$n = 200$, $n_2 = 5$, $S = 10,000$.

δ	LC	P=1	P=2	P=3	P=4	P=5
0.0	0.963	0.156	0.605	0.851	0.918	0.949
0.2	0.885	0.964	0.076	0.149	0.318	0.535
0.4	0.792	1.000	0.030	0.038	0.062	0.126

$n = 200$, $n_2 = 25$, $S = 10,000$.

δ	LC	P=1	P=2	P=3	P=4	P=5
0.0	0.990	0.251	0.855	0.943	0.960	0.970
0.2	0.945	0.979	0.125	0.259	0.490	0.681
0.4	0.893	1.000	0.024	0.045	0.081	0.149

$n = 200$, $n_2 = 50$, $S = 10,000$.

δ	LC	P=1	P=2	P=3	P=4	P=5
0.0	0.988	0.279	0.918	0.968	0.985	0.989
0.2	0.964	0.989	0.097	0.247	0.497	0.766
0.4	0.916	1.000	0.015	0.032	0.075	0.177

$n = 200$, $n_2 = 100$, $S = 10,000$.

δ	LC	P=1	P=2	P=3	P=4	P=5
0.0	0.982	0.196	0.902	0.951	0.969	0.983
0.2	0.949	0.944	0.064	0.282	0.682	0.885
0.4	0.889	0.998	0.001	0.004	0.026	0.211

$n = 200$, $n_2 = 150$, $S = 10,000$.

δ	LC	P=1	P=2	P=3	P=4	P=5
0.0	0.963	0.076	0.844	0.916	0.958	0.981
0.2	0.903	0.706	0.083	0.557	0.846	0.948
0.4	0.772	0.945	0.000	0.000	0.142	0.757

may be advisable to increase S to confirm that this is not simply a consequence of too few splits of the data being considered. Our experience with this approach is that $S = 10,000$ is sufficient to overcome such concerns.

3.2. **Finite-Sample Performance: Time-Series Data.** The time-series literature dealing with predictive accuracy and forecasting is quite vast, and we make no claims at surveying this literature here.[13] Early work on forecast model comparison by Ashley, Granger & Schmalensee (1980) and Granger & Newbold (1986) generated broad interest in this topic. However, only recently have formal tests that directly relate to forecast accuracy and predictive ability surfaced. Most notably the available tests include Diebold & Mariano (1995) (the 'DM' test) and its size corrected counterpart

[13]See the review by De Gooijer & Hyndman (2006) for a thorough and up-to-date survey and bibliography on the subject.

Harvey, Leybourne & Newbold (1997) (the 'MDM' test) along with those proposed by Swanson & White (1997), Ashley (1998), Harvey, Leybourne & Newbold (1998), West & McCracken (1998), Harvey & Newbold (2000), Corradi & Swanson (2002), van Dijk & Franses (2003), Hyndman & Koehler (2006) and Clark & West (2007), among others. Given the popularity of Diebold & Mariano's (1995) test[14] we perform a simple Monte Carlo simulation similar to that presented in Section 3 but with stationary time-series models as opposed to cross-section ones.

We generate data from an AR(2) model given by

$$y_t = \rho_1 y_{t-1} + \rho_2 y_{t-2} + \varepsilon_t, \tag{8}$$

where $\rho_1 = 0.9$ for all simulations, ρ_2 varies from 0 (an AR(1) model) to -0.8 in increments of 0.2 and ε_t is $N(0,1)$. For all simulations we conduct $M = 1,000$ Monte Carlo replications using $S = 10,000$ sample splits for our revealed performance approach. We use sample sizes of $n = 200$ and $n = 400$ holding out the last $n_2 = 5, 10, 25$, or 50 observations of each resample with which to forecast.[15] When $\rho_2 = 0$ we can determine the extent to which our test predicts (less than or) equivalent accuracy of the forecasts, while when $\rho_2 \neq 0$ we can assess how often our method determines that an AR(2) model predicts better than an AR(1) when indeed it should. We also compare the AR(1) to MA(1) and MA(2) specifications by way of comparison. We compare our results with the DM and MDM test, noting that the DM test has a tendency to over-reject hence our inclusion of the size-corrected MDM results. We report empirical rejection frequencies at the 5% level in Tables 3 and 4.

First, a comparison of the DM and MDM results indicate that indeed the DM test suffers from rather substantial upwards size distortions. We direct the interested reader to Harvey et al. (1998)[16] who report on the formal size and power properties of the DM and MDM tests. Comparing Table 3 ($n = 200$) with 4 ($n = 400$) we see that the upwards size distortion (i.e., $\rho_2 = 0$, test AR(1) versus AR(2)) falls as n increases. In fact, the DM test's size distortions are so large as to all but remove it from contention, at least from this simulation. The size corrected MDM test, however, performs well, at least from the perspective of correcting for upward size distortions, though again there appears to be some minor distortion (i.e., $\rho_2 = 0$, test AR(1) versus AR(2)). All three approaches increase in power as n_2 increases as expected, however, the RP test approaches 1 faster suggesting that smaller holdout samples are required to acquire knowledge of predictive accuracy. This aspect of our approach overcomes one known drawback of the MDM and related tests, namely, the need to have a sufficiently large hold-out sample in order for the test to have power. Lastly, comparing the RP and MDM test one will see immediately that as ρ_2 moves away from zero our rejection

[14]Our use of 'popularity' stems from the fact that this paper was named one of the 10 best papers in the 20$^{\text{th}}$ anniversary edition of the *Journal of Business & Economic Statistics* and a Social Science Citations Index search, conducted on December 4, 2008, of this paper yielded 500 citations.

[15]For each Monte Carlo simulation, the initial data generated is passed through the automatic block length selection mechanism of Politis & White (2004) to determine the optimal block length. This block length is then used for each of the S splits of the data.

[16]See also Harvey & Newbold (2000), Meade (2002) and van Dijk & Franses (2003).

TABLE 3. Each entry represents rejection frequencies 5% level for the test that the AR(1) model has predictive accuracy equal to that for each model in the respective column heading, rejecting when the model in the respective column heading has improved predictive accuracy.

$n = 200$, $n_2 = 5$, $S = 10,000$.

	DM Test			MDM Test			RP Test		
ρ_2	AR(2)	MA(1)	MA(2)	AR(2)	MA(1)	MA(2)	AR(2)	MA(1)	MA(2)
0.0	0.471	0.205	0.205	0.031	0.003	0.003	0.000	0.000	0.000
-0.4	0.658	0.535	0.535	0.119	0.019	0.019	0.846	0.322	0.322
-0.8	0.896	0.941	0.941	0.259	0.259	0.259	0.993	0.993	0.993

$n = 200$, $n_2 = 10$, $S = 10,000$.

	DM Test			MDM Test			RP Test		
ρ_2	AR(2)	MA(1)	MA(2)	AR(2)	MA(1)	MA(2)	AR(2)	MA(1)	MA(2)
0.0	0.460	0.130	0.130	0.045	0.005	0.005	0.001	0.000	0.000
-0.4	0.715	0.533	0.533	0.171	0.041	0.041	0.847	0.329	0.329
-0.8	0.959	0.932	0.932	0.562	0.630	0.630	1.000	1.000	1.000

$n = 200$, $n_2 = 25$, $S = 10,000$.

	DM Test			MDM Test			RP Test		
ρ_2	AR(2)	MA(1)	MA(2)	AR(2)	MA(1)	MA(2)	AR(2)	MA(1)	MA(2)
0.0	0.447	0.026	0.026	0.043	0.002	0.002	0.003	0.000	0.000
-0.4	0.822	0.564	0.564	0.270	0.066	0.066	0.895	0.388	0.388
-0.8	0.988	0.972	0.972	0.818	0.771	0.771	0.986	0.986	0.986

$n = 200$, $n_2 = 50$, $S = 10,000$.

	DM Test			MDM Test			RP Test		
ρ_2	AR(2)	MA(1)	MA(2)	AR(2)	MA(1)	MA(2)	AR(2)	MA(1)	MA(2)
0.0	0.396	0.002	0.002	0.048	0.000	0.000	0.008	0.000	0.000
-0.4	0.906	0.573	0.573	0.411	0.070	0.070	0.881	0.388	0.388
-0.8	0.999	0.981	0.981	0.959	0.898	0.898	0.993	0.993	0.993

frequencies approach one at a faster rate that the MDM test, indicating that this approach is capable of detecting gains in predictive accuracy outside of an *iid* setting.

We note that the ability to choose one's loss function when using our approach may be appealing to practitioners. For instance, if interest lies in penalizing more heavily over or underprediction, the use of asymmetric loss functions may be of interest (LINEX for example, Chang & Hung 2007). See Efron (1983, 1986) for more on issues related to prediction rules and apparent error in relation to cross-validation and bootstrapping.

4. Empirical Illustrations

4.1. **Application to Wooldridge's wage1 Data.** For what follows, we consider an application that involves multiple regression analysis with qualitative information. This example is taken from Wooldridge (2003, pg. 226).

TABLE 4. Each entry represents rejection frequencies 5% level for the test that the AR(1) model has predictive accuracy equal to that for each model in the respective column heading, rejecting when the model in the respective column heading has improved predictive accuracy.

$n = 400$, $n_2 = 5$, $S = 10,000$.

	DM Test			MDM Test			RP Test		
ρ_2	AR(2)	MA(1)	MA(2)	AR(2)	MA(1)	MA(2)	AR(2)	MA(1)	MA(2)
0.0	0.510	0.204	0.204	0.031	0.001	0.001	0.000	0.000	0.000
-0.4	0.702	0.552	0.552	0.126	0.020	0.020	0.811	0.279	0.279
-0.8	0.910	0.953	0.953	0.356	0.000	0.000	0.995	0.991	0.991

$n = 400$, $n_2 = 10$, $S = 10,000$.

	DM Test			MDM Test			RP Test		
ρ_2	AR(2)	MA(1)	MA(2)	AR(2)	MA(1)	MA(2)	AR(2)	MA(1)	MA(2)
0.0	0.434	0.081	0.081	0.022	0.007	0.007	0.000	0.000	0.000
-0.4	0.737	0.564	0.564	0.158	0.060	0.060	0.890	0.351	0.351
-0.8	0.934	0.948	0.948	0.556	0.588	0.588	0.991	0.991	0.991

$n = 400$, $n_2 = 25$, $S = 10,000$.

	DM Test			MDM Test			RP Test		
ρ_2	AR(2)	MA(1)	MA(2)	AR(2)	MA(1)	MA(2)	AR(2)	MA(1)	MA(2)
0.0	0.521	0.043	0.043	0.043	0.000	0.000	0.007	0.000	0.000
-0.4	0.835	0.558	0.558	0.273	0.053	0.053	0.901	0.411	0.411
-0.8	0.988	0.976	0.976	0.800	0.778	0.778	0.993	0.993	0.993

$n = 400$, $n_2 = 50$, $S = 10,000$.

	DM Test			MDM Test			RP Test		
ρ_2	AR(2)	MA(1)	MA(2)	AR(2)	MA(1)	MA(2)	AR(2)	MA(1)	MA(2)
0.0	0.351	0.000	0.000	0.047	0.000	0.000	0.000	0.000	0.000
-0.4	0.907	0.579	0.579	0.410	0.059	0.059	0.909	0.432	0.432
-0.8	1.000	0.992	0.992	0.944	0.919	0.919	0.997	0.997	0.997

We consider modelling an hourly wage equation for which the dependent variable is log(wage) (lwage) while the explanatory variables include three continuous variables, namely educ (years of education), exper (the number of years of potential experience), and tenure (the number of years with their current employer) along with two qualitative variables, female ('Female'/'Male') and married ('Married'/'Notmarried'). For this example there are $n = 526$ observations. We use Hurvich, Simonoff & Tsai's (1998) AIC_c approach for bandwidth selection.

We first test a parametric model that is linear in all variables for correct parametric specification using Ramsey's (1969) RESET test for functional form. We obtain a P-value of 0.0005104 and reject the null of correct specification. We then estimate a nonparametric (local linear) model and test the null that the nonparametric model and a parametric model that is linear in all variables have equal ASPE versus the alternative that the parametric model has greater ASPE. This yields a P-value of $6.969649e-50$ hence we reject the null and conclude that the nonparametric model has statistically significantly improved performance on independent data and therefore represents

a statistical improvement over the rejected parametric specification. If we consider a model that is popular in the literature (quadratic in experience) we again reject the null that the model is correctly specified based on the RESET test (P-value of 0.0009729) and again reject the null that the parametric and nonparametric specifications are equivalent in terms of their ASPE and conclude that the nonparametric specification is preferred (the P-value is $9.416807e-05$). These results indicate that the proposed methodology can be successfully used in applied settings.

Figures 1 and 2 present boxplots and empirical distribution functions for the vector of length S of ASPEs for each model along with median and mean values for each.[17] It can be seen from Figure 2 that a stochastic dominance relationship exists again indicating that the nonparametric model is to be preferred on the basis of its performance on independent data.

4.1.1. *Implications of Non-Optimal Smoothing.* It is surprisingly common to encounter practitioners applying kernel methods using non-optimal rules of thumb for bandwidth choice, and in this section we briefly examine the issue of non-optimal smoothing for the method proposed herein. We consider the two parametric models considered above, namely, one that is linear in all variables and one reflecting the popular specification that is quadratic in experience. We report three nonparametric models, one that is optimally smoothed according to Hurvich et al.'s (1998) AIC_c criterion, one that is undersmoothed (25% of the bandwidth values given by the AIC_c criterion), and one that is oversmoothed using the maximum possible bandwidths for all variables (0.5 for the discrete variables and ∞ for the continuous ones). We report the apparent error given by $\text{ASE} = n^{-1}\sum_{i=1}^{n}(y_i^* - \hat{g}_{z^n}(z_i^*))^2$, and the mean estimated true error taken over all sample splits, $S^{-1}\sum_{j=1}^{S}\ \text{APSE}_j$ where $\text{ASPE}_j = n_2^{-1}\sum_{i=n_1+1}^{n}(y_i^* - \hat{g}_{z^{n_1}}(z_i^*))^2$, $j = 1, 2, \ldots, S$. Results are reported in Table 5.

TABLE 5. Estimated apparent and true errors for the nonparametric and parametric models for Wooldridge's wage1 data.

Nonparametric Model		
Smoothing	Apparent Error	True Error
Undersmoothed	0.1193826	0.1685334
AIC_c	0.1371530	0.1605222
Oversmoothed	0.1928813	0.1679134

Parametric Model		
Experience	Apparent Error	True Error
Linear	0.1681791	0.1723598
Quadratic	0.1590070	0.1634519

[17]A box-and-whisker plot (sometimes called simply a 'box plot') is a histogram-like method of displaying data, invented by J. Tukey. To create a box-and-whisker plot, draw a box with ends at the quartiles Q_1 and Q_3. Draw the statistical median M as a horizontal line in the box. Now extend the 'whiskers' to the farthest points that are not outliers (i.e., that are within 3/2 times the interquartile range of Q_1 and Q_3). Then, for every point more than 3/2 times the interquartile range from the end of a box, draw a dot.

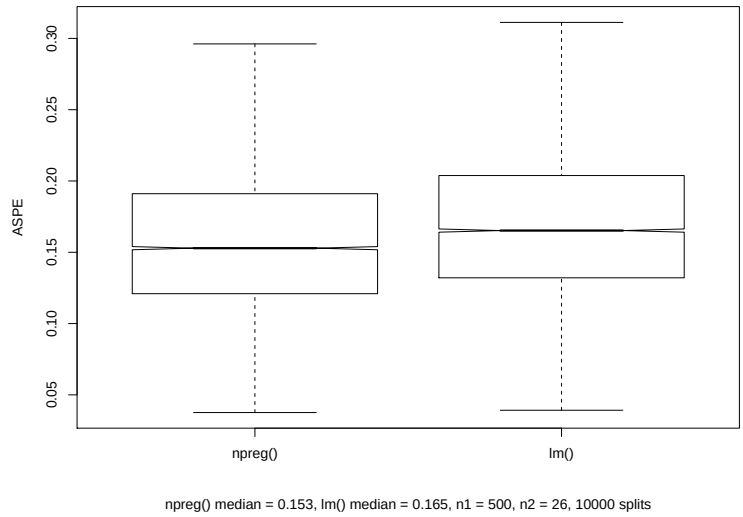

FIGURE 1. Boxplots of the ASPE for the $S = 10,000$ splits of the data for the wage1 dataset. Median values for each model appear in the subtitle below the figure.

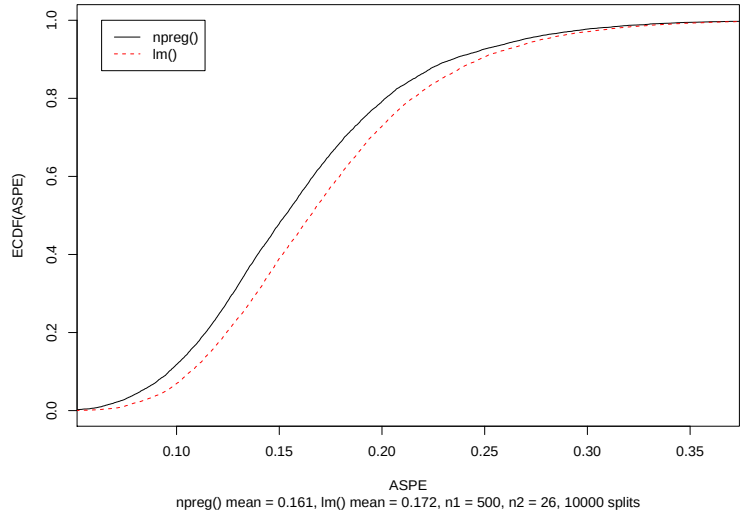

FIGURE 2. Empirical distribution functions of the ASPE for the $S = 10,000$ splits of the data for the wage1 dataset. Mean values for each model appear in the subtitle below the figure.

It can be seen from Table 5 that the undersmoothed and optimally smoothed apparent errors are indeed overly optimistic as are those for the linear and quadratic parametric models, as expected. Interestingly, the oversmoothed nonparametric model is overly pessimistic. The tests provided in

Section 4.1 above are tests that the value in column 3 of Table 5 for the AIC_c model (0.1605222) is statistically significantly lower than that for the values in column 3 for both the linear (0.1723598) and quadratic (0.1634519) models. Thus, the nonparametric model is 7.4% more efficient than the linear model and 1.8% more efficient that the quadratic model as measured in terms of performance on independent data while the quadratic model is 5.5% more efficient than the linear model.

4.2. **Application to CPS Data.** We consider a classic data set taken from Pagan & Ullah (1999, page 155) who consider Canadian cross-section wage data consisting of a random sample obtained from the 1971 Canadian Census Public Use (CPS) Tapes for male individuals having common education (Grade 13). There are $n = 205$ observations in total, and 2 variables, the logarithm of the individuals wage (logwage) and their age (age). The traditional wage equation is typically modelled as a quadratic in age.

For what follows we consider parametric models of the form

$$\log(\text{wage})_i = \beta_0 + \sum_{j=1}^{P} \beta_j \text{age}_i^j + \varepsilon_i$$

When $P = 1$ we have a simple linear model, $P = 2$ quadratic and so forth. These types of models are ubiquitous in applied data analysis

For each model we apply the RESET test. Table 6 summarizes the model specification tests for $P = 1$ through $P = 7$.

TABLE 6. Ramsey's (1969) RESET test for correct specification of the parametric models for the Canadian CPS data.

P	RESET	df1	df2	p-value
1	26.2554	2	201	7.406e-11
2	13.1217	2	200	4.42e-06
3	11.34	2	199	2.168e-05
4	2.1999	2	198	0.1135
5	0.8488	2	197	0.4295
6	1.0656	2	196	0.3465
7	1.4937	2	195	0.2271

Models with $P > 3$ pass this specification test. However, this does not imply that the model will outperform other models on independent data drawn from this DGP. The model may be overspecified, and may also reflect issues regarding the RESET test's power.

We now consider applying the proposed method to this dataset, considering parametric models of order $P = 1$ through 7 along with the local constant and linear nonparametric specifications. We present results in the form of boxplots and empirical cumulative distributions in Figures 3 and 4. The boxplots and ECDFs for $P = 4, 5$ or 6 reveal that these models exhibit visual stochastic dominance relationships with the parametric models for $P = 1, 2, 3$ and 7. This is suggestive that the models $P = 1, 2, 3$ may be underspecified while the model $P = 7$ is perhaps overspecified.

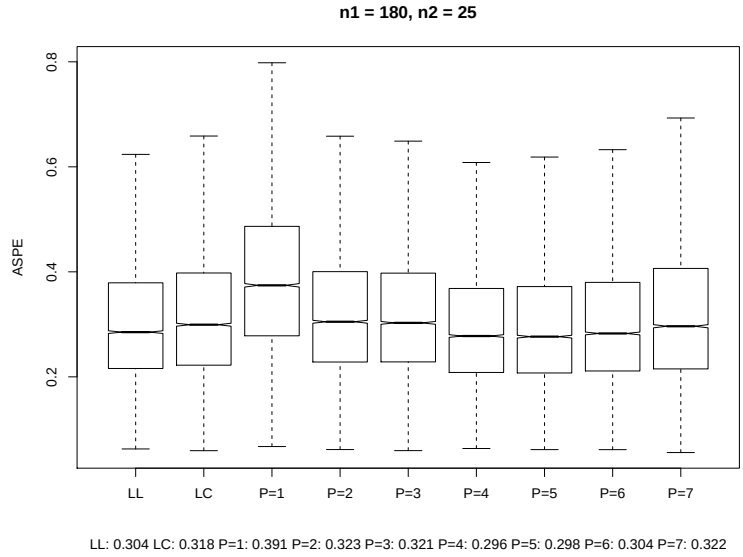

FIGURE 3. Boxplots of the ASPE for the $S = 10,000$ splits of the Canadian CPS data. Median values for each model appear in the subtitle below the figure.

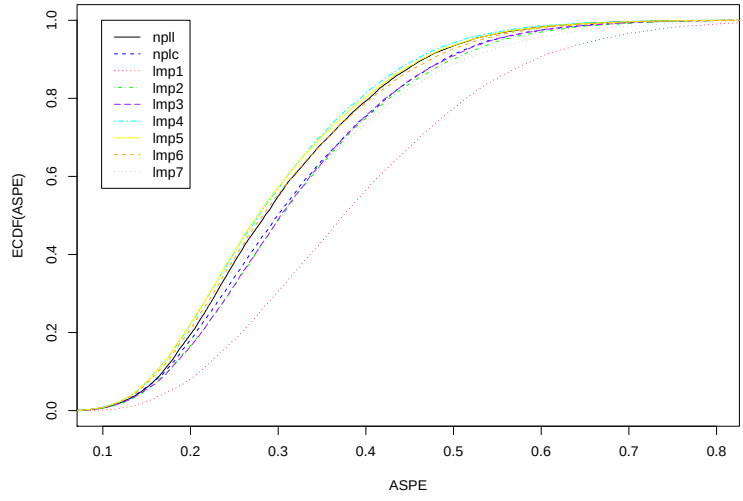

FIGURE 4. Empirical distribution functions of the ASPE for the $S = 10,000$ splits of the Canadian CPS data.

The t-statistics and P-values for the test that the mean ASPE is equal for each model versus the local linear model are given in Table 7.

Table 7 reveals that the local linear specification is preferred to the local constant specification on true error grounds. Furthermore, the popular linear and quadratic specifications are dominated

TABLE 7. RP test results for the Canadian CPS data. Small P-values indicate that the nonparametric local linear model performs better than the model listed in column 1 according to the true error criterion.

Model	t	P-value
LC	7.847834	2.222817e-15
$P = 1$	45.70491	0
$P = 2$	10.85293	1.152248e-27
$P = 3$	9.682618	1.999341e-22
$P = 4$	-4.796251	0.9999992
$P = 5$	-3.810738	0.9999305
$P = 6$	-0.2022363	0.580133
$P = 7$	9.840635	4.257431e-23

by the local linear specification as is the less common cubic specification. The quartic and quintic parametric specifications dominate the local linear specification as would be expected given the findings of Murphy & Welch (1990). Interestingly, the local linear specification dominates the overspecified parametric model (P= 7), again underscoring the utility of the proposed approach.

4.3. **Application to Housing Data.** Hedonic analysis of housing data was studied in Anglin & Gençay (1996).[18] They argued that standard parametric models, which passed the RESET test, were outperformed based on overall fit against a partially linear model; two different tests of linearity versus a partially linear model rejected the null hypothesis of correct linear specification. Moreover, to further emphasize the superior performance of the partially linear model, they conducted two separate sampling exercises. First, they looked at price predictions for a 'reference' house and plotted the change in price of the home as the number of bedrooms changed. Their results suggested that the price predictions from the semiparametric model were statistically different at the 99% level from the parametric predictions and the parametric model had wider confidence bounds than the partially linear model. Second, Anglin & Gençay (1996) performed a similar hold-out sample exercise as discussed here, however, they do not repeat this exercise a large number of times. From their paper it appears that they did this for *one* sampling of the data using first 10 hold-out homes and then using 20 hold-out homes. The holdout homes were randomly selected.

Recently, Parmeter, Henderson & Kumbhakar (2007) challenged the partially linear specification of Anglin & Gençay (1996) and advocated for a fully nonparametric approach. Their findings suggested that the partially linear model fails to pass a test of correct specification against a nonparametric alternative, that Anglin & Gençay's (1996) measure of within-sample fit of the partially linear model was overstated, and the inclusion of categorical variables as continuous variables into the unknown function may produce a loss of efficiency Racine & Long (2008). This collection of

[18]The data from their paper is available on the JAE data archives webpage or can be found in the Ecdat package (Croissant 2006) in R (R Development Core Team 2008) under the name 'housing'.

results provides a useful conduit for examining the revealed performance of the parametric specification of Anglin & Gençay (1996, Table III), the *appropriate*[19] partially linear specification of Anglin & Gençay (1996), and the fully nonparametric specification of Parmeter et al. (2007).

Formally, we model a hedonic price equation where our dependent variable is the logarithm of the sale price of the house while the explanatory variables include six categorical variables, namely if the house is located in a preferential area in the Windsor, Canada area and if the house has air conditioning, gas heated water, a fully finished basement, a recreational room, and a driveway, four ordered covariates, the number of garage places, the number of bedrooms, full bathrooms and stories of the house, and a single continuous variable, the logarithm of the lot size of the house ($\ln(lot)$). There are a total of $n = 546$ observations for this data. All bandwidths are selected using the AIC_c criterion.[20]

Our three models are:

$$\ln(sell) = \gamma_{cat} z_{cat} + \gamma_{ord} z_{ord} + \beta \ln(lot) + \varepsilon_1 \tag{9}$$

$$\ln(sell) = \gamma_{cat} z_{cat} + g_{AG}(z_{ord}, \ln(lot)) + \varepsilon_2 \tag{10}$$

$$\ln(sell) = g_{PHK}(z_{cat}, z_{ord}, \ln(lot)) + \varepsilon_3, \tag{11}$$

where z_{cat} and z_{ord} are the vectors of categorical and ordered variables, respectively, described above.[21] We denote the unknown functions in equations (10) and (11) by AG and PHK to refer to the models in Anglin & Gençay (1996) and Parmeter et al. (2007). As noted by Anglin & Gençay (1996), the parametric model is not rejected by a RESET test, suggesting correct specification.[22]

Our test of revealed performance begins with the estimation of all three models and then tests three distinct null hypotheses. First, we test if the nonparametric and linear models have equal ASPE, second we test if the nonparametric and partially linear models have equal ASPE and thirdly, we test if the linear and partially linear models have equal ASPE. For all three tests our alternative hypothesis is that the less general model has a greater ASPE. These tests yield P-values of 1, $2.2e-16$ and 1, suggesting that the linear model has superior predictive performance over both the appropriately estimated semiparametric model of Anglin & Gençay (1996) and the fully nonparametric model of Parmeter et al. (2007), while the fully nonparametric model has performance that is at least as good as the semiparametric model. This is in direct contrast to Anglin & Gençay's (1996) finding that the semiparametric model provides lower MPSEs for

[19]We do not estimate the partially linear model as it appears in Anglin & Gençay (1996) since Parmeter et al. (2007) were unable to exactly replicate their results and Anglin & Gençay's (1996) handling of ordered discrete variables as continuous is erroneous given the current practice of using generalized kernel estimation.

[20]See Racine & Long (2008) for more on bandwidth selection in partially linear models.

[21]We note that although the number of garage places is an ordered variable, Anglin & Gençay (1996) did not include it in the unknown function in their partially linear setup. To be consistent with their modelling approach we follow suit and have the number of garage places enter in the linear portion of (10).

[22]Anglin & Gençay (1996, pg. 638) do note, however, that their benchmark model is rejected using the specification test of Wooldridge (1992). Also, we use the model estimated in Table III of Anglin & Gençay (1996) since this model has a higher $\bar{R}^2$ and as they note (Anglin & Gençay 1996, page 638) the performance of this model is not substantially different from their benchmark model.

their holdout samples and is an immediate consequence of the fact that they did not repeat their sampling process a large number of times. Additionally, Gençay & Yang (1996) and Bin (2004), also in a hedonic setting, compare semiparametric out-of-sample fits against parametric counterparts using only *one* sample. These setups are entirely incorrect for assessing if one model produces substantially better out-of-sample predictions than another.

Figures 5 and 6 present boxplots and empirical distribution functions for the vector of length S of ASPEs for each of the three models along with median and mean values for each. It can be seen from Figure 6 that a stochastic dominance relationship exists between the linear model (lm()) and both the nonparametric and partially linear models (npreg() and npplreg()), again indicating that the linear model is to be preferred on the basis of its performance on independent data. Figure 5 is not suggestive of a stochastic dominance relationship between the linear model and the nonparametric model whereas the plots of the ECDFS in Figure 6 readily reveal that the parametric model dominates both the nonparametric and partly linear model, suggesting the use of both plots when assessing the performance of two competing models.

4.4. **Application to Economic Growth Data.** Recent studies by Maasoumi, Racine & Stengos (2007) and Henderson, Papageorgiou & Parmeter (2008) have focused on fully nonparametric estimation of 'Barro regressions' (see Durlauf, Johnson & Temple 2005) and argue persuasively that standard linear models of economic growth cannot adequately capture the nonlinearities that are most likely present in the underlying growth process. While both papers have soundly rejected basic linear specifications as well as several sophisticated parametric models, it is not evident that the nonparametric model explains the growth data any better than a parametric model.

For this example we use the dataset 'oecdpanel' available in the np package (Hayfield & Racine 2008) in R (R Development Core Team 2008). This panel covers seven 5 year intervals beginning in 1960 for 86 countries.[23] Our dependent variable is the growth rate of real GDP per capita over each of the 5 year intervals while our explanatory variables include an indicator if the country belongs to the OECD, an ordered variable indicating the year, and the traditional, continuous 'Solow' variables: the initial real per capita GDP, the average annual population growth over the 5 year interval, the average investment-GDP ratio over the 5 year period and the average secondary school enrollment rate for the 5 year period. This is the same data used in Liu & Stengos (1999) and Maasoumi et al. (2007).

We compare the baseline, linear specification and a model that includes higher order terms in initial GDP and human capital to a local linear nonparametric model with bandwidths selected via AIC_c. The baseline linear model is rejected for correct specification using a RESET test (P-value=0.03983), but the higher order model cannot be rejected using a RESET test (P-value=0.1551). However, this test result could be due to power problems related to overspecification with the inclusion of the additional quadratic, cubic and quartic terms. Using the consistent model specification test of Hsiao, Li & Racine (2007), the higher order parametric model is rejected

[23]See Liu & Stengos (1999, Table 1) for a list of countries in the dataset.

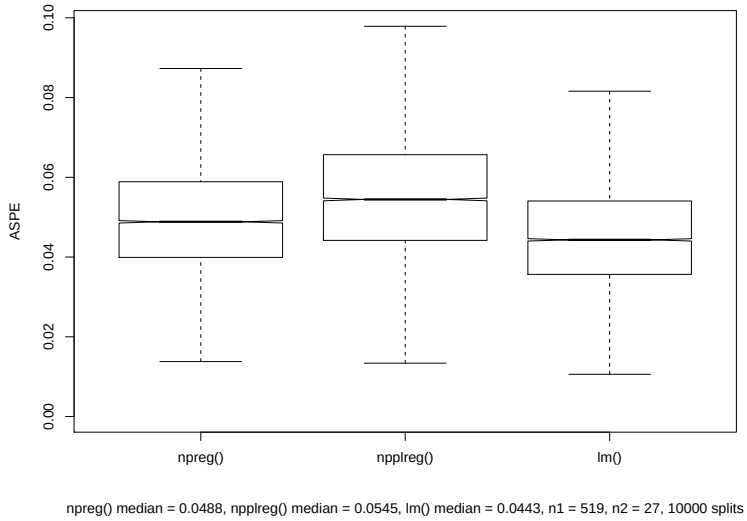

FIGURE 5. Boxplots of the ASPE for the $S = 10,000$ splits of the housing data. Median values for each model appear in the subtitle below the figure.

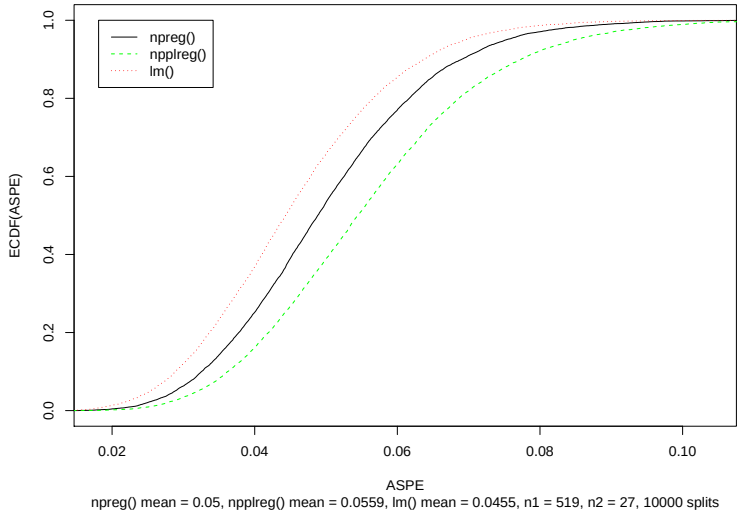

FIGURE 6. Empirical distribution functions of the ASPE for the $S = 10,000$ splits of the housing data. Mean values for each model appear in the subtitle below the figure.

for correct specification with a P-value of 4.07087e-06. The question remains however, does the nonparametric model predict growth any better than this higher order model?

Our test results, boxplots, and empirical CDFs all suggest that the nonparametric model significantly outperforms both the baseline 'Barro' regression model and the higher order parametric model presented in Maasoumi et al. (2007). Our P-values for tests of equality between either the baseline linear model or the higher order linear model and the local linear nonparametric model are $3.475388e-06$ and $1.542881e-07$, respectively. This is suggestive that neither parametric model is revealing superior performance to the nonparametric model which corroborates the findings of Maasoumi et al. (2007) and Henderson et al. (2008).

Figures 7 and 8 present boxplots and empirical distribution functions for the vector of length S of ASPEs for each of the three models along with median and mean values for each. It can be seen from Figure 8 that a stochastic dominance relationship exists between the nonparametric model (npreg()) and both of the linear models (lm() and ho-lm()), again indicating that the nonparametric model is to be preferred on the basis of its performance on independent draws from the data. What is interesting is that in terms of revealed performance given through the ECDFs, the higher order linear model *does not* exhibit any stochastic dominance over the standard 'Barro' regression, suggesting that the hypothesized nonlinearities present are more complex that simple power terms of the individual covariates. Interestingly, Henderson et al. (2008) have uncovered marginal effects of the covariates consistent more so with interactions between covariates than higher order terms of individual covariates.

4.5. **Application to Interest Rate Data.** In this section we consider modelling the U.S. federal funds interest rate. The data is taken from Davidson & MacKinnon (2004, page 601). This data is a time-series of monthly observations on the interest rate from January 1955 to December 2001 (a total of $n = 564$ observations). Figures 9 and 10 display the raw and first differenced interest rate series as well as auto-covariance and partial autocorrelation functions for 20 lags.

Before proceeding we note that we reject stationarity at the 5% level for the raw series, hence we first difference our data at which point we no longer reject stationarity. Next, we note that the the large spike of our differenced data for the first autocorrelation (Figure 10, lower right panel) suggests that a MA(1) process may be present. However, the presence of positive and significant autocorrelations past lag 1 and the regular pattern in the distant autocorrelations suggests that a more complex data generating process may be present. Also, the partial autocorrelations (Figure 10, lower left panel), which have a positive and significant spike for lag 1 and a negative and significant spike for lag 2 would rule out the use of an AR(1) model but could be consistent with and AR(2) model.[24] Fitting the best ARIMA process to the first differenced data suggests that an MA(1) process is appropriate.[25] For our empirical exercise we use a holdout sample of 12 observations, which corresponds to one year for the undifferenced data, for our proposed test. The automatic block length selection of Politis & White (2004) suggests we use a random block length of 4 when resampling.

[24]The positive and significant partial autocorrelations at lag 8, 13, and 26 are difficult to interpret.

[25]This was done using the the entire dataset with the auto.arima() function in the forecast package (Hyndman 2008) in R (R Development Core Team 2008).

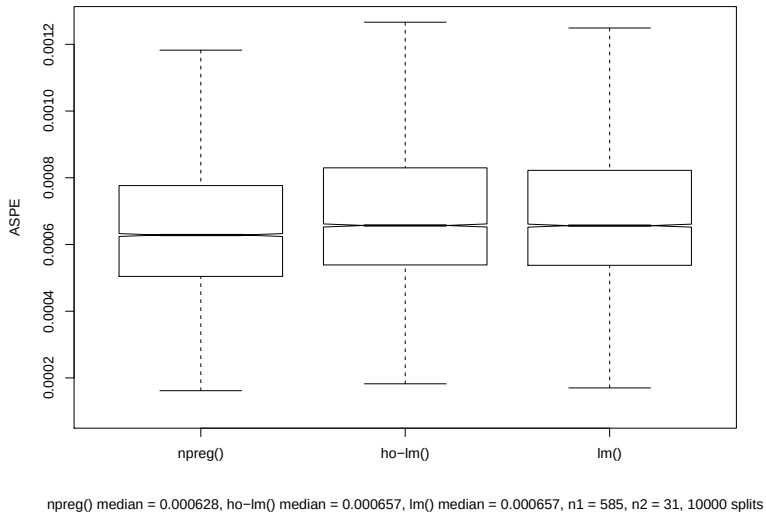

FIGURE 7. Boxplots of the ASPE for the $S = 10,000$ splits of the oecdpanel data. Median values for each model appear in the subtitle below the figure.

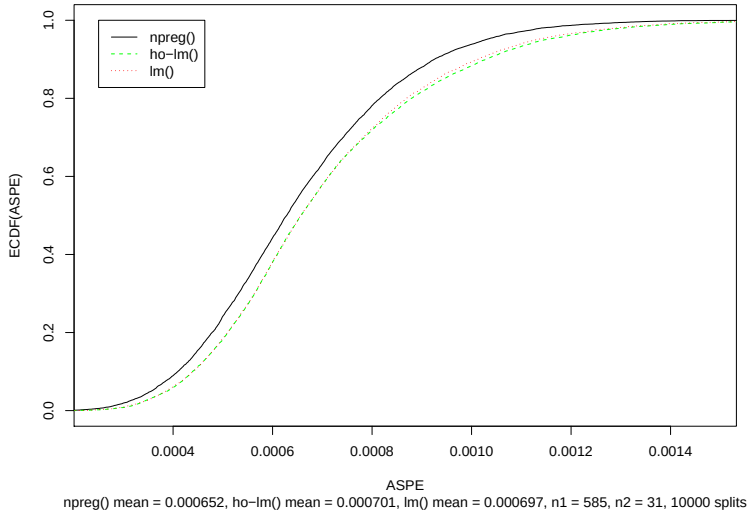

FIGURE 8. Empirical distribution functions of the ASPE for the $S = 10,000$ splits of the oecdpanel data. Mean values for each model appear in the subtitle below the figure.

Before providing our graphical evidence we compare our 12 1-step ahead forecasts from an AR(1), AR(2), MA(1), ARMA(1,1), ARMA(2,1) and an ARMA(2,2) model with the MA(1) being our baseline model. The $p-$values of our test, along with those of the DM and the MDM test appear

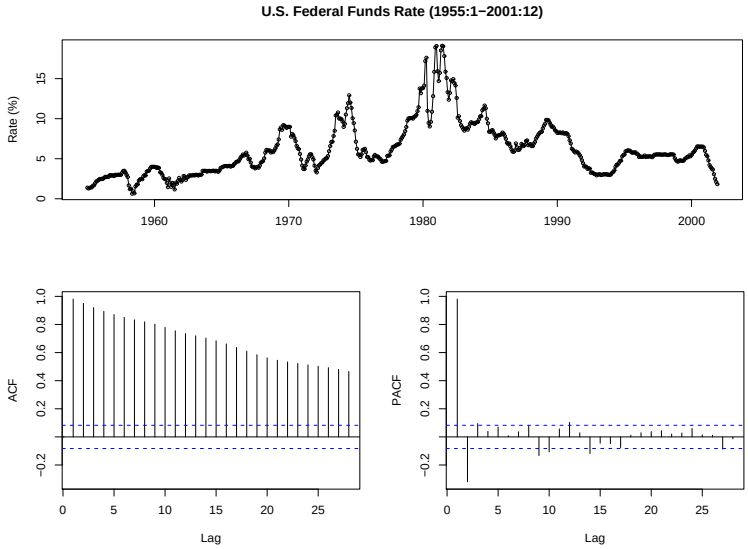

FIGURE 9. Time plot, auto-covariance and partial autocorrelation plots for the federal funds interest rate.

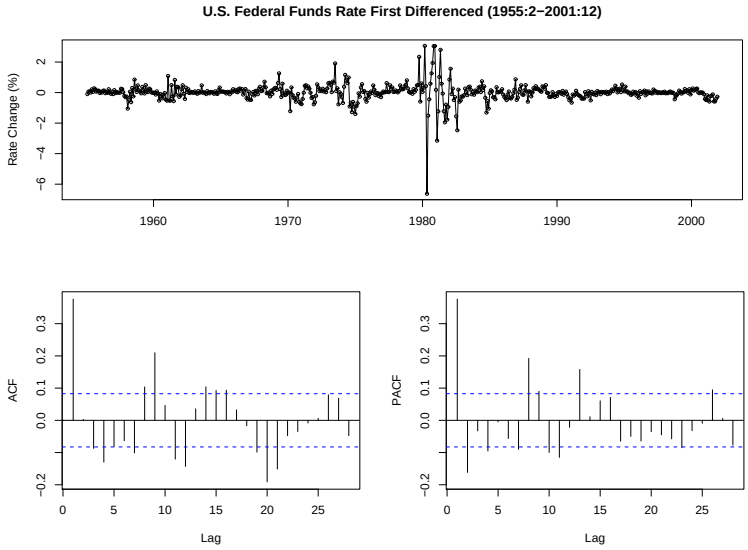

FIGURE 10. Time plot, auto-covariance and partial autocorrelation plots for the first differenced federal funds interest rate time-series.

in Table 8. While all three tests suggest that we cannot reject the null of equal forecast accuracy for the AR(1) and ARMA(1,1) against the MA(1) at all conventional significance levels, both the DM and MDM tests reject the null hypothesis of equal forecast accuracy for the ARMA(2,1) and ARMA(2,2) at the 1% level of significance against the MA(1) and reject the AR(2) against

the MA(1) at the 10% level. Our test reveals that we cannot reject the null of equal revealed performance (forecast accuracy in this setting) for all five of the alternative models relative to the MA(1) process.

TABLE 8. DM, MDM, and RP test results for the first differenced federal funds rate time-series. Our null hypothesis is that the model of interest has equal 1-step ahead forecasts under squared error loss relative to a MA(1) process.

Model	DM	MDM	RP
AR(1)	1	0.9560	0.4185
AR(2)	0.0324	0.0520	0.4240
ARMA(1,1)	1	0.9998	0.4344
ARMA(2,1)	1.570e-04	0.0027	0.4239
ARMA(2,2)	6.885e-05	0.0019	0.4762

Our boxplot and ECDF comparison plots appear in Figures 11 and 12. Unlike our cross-sectional results, stochastic dominance relationships are not visually present. One must study the plots carefully to detect visual discrepancies in both the ECDFs and the boxplots for all six models. The AR(2) has the lowest upper quartile for APSE of the six models but overall these six models appear to be have distributions of expected true errors that appear equal over 10,000 bootstrap simulations. Thus, our RP test results confirm the visual evidence.

While our demonstration of the RP test for time-series data has not been as exhaustive as our cross sectional applications, we present this illustrative application simply to demonstrate that this method can be used in time-series settings when one uses an appropriate bootstrap procedure. A more detailed analysis of the performance of this test in dependent data settings would be a fruitful avenue for future investigation.

5. Conclusion

In this paper we propose a general methodology for assessing the predictive performance of competing approximate models based on resampling ideas. Our idea is to take repeated hold-out samples (appropriately constructed) from the data at hand to create an estimate of the expected true error of a model. A model with a lower expected true error predicts better *on* average than a model with a higher expected true error and can therefore be expected to be closer to the underlying data generating process. Our approach allows practitioners to compare a broad range of modelling alternatives and is not limited to the regression-based examples provided herein. Overall, this method can be used to determine whether or not a more flexible model offers any gains in terms of expected performance than a less complex model and provides an alternative avenue for direct comparison of parametric and nonparametric regression surfaces (e.g., Härdle & Marron 1990, Härdle & Mammen 1993).

We have presented both simulated and empirical evidence underscoring the utility of the proposed method for both *iid* and dependent data settings. Our simulation results indicate that, relative

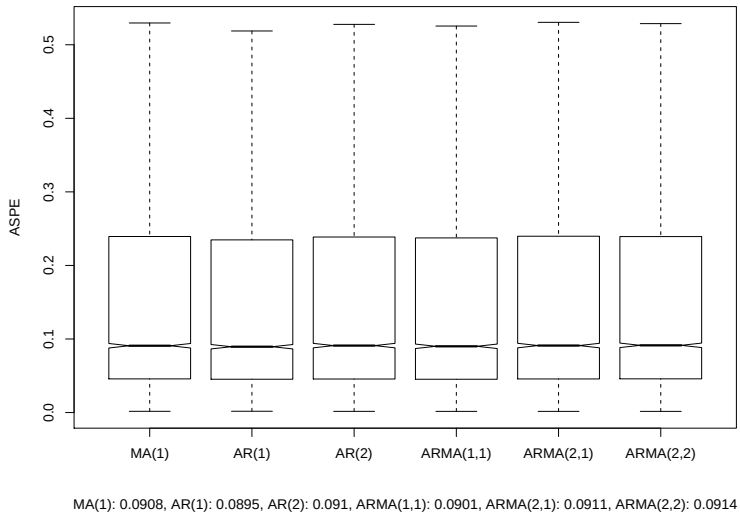

FIGURE 11. Boxplots of the ASPE for the $S = 10,000$ splits of the first differenced federal funds rate time-series data. Median values for each model appear in the subtitle below the figure.

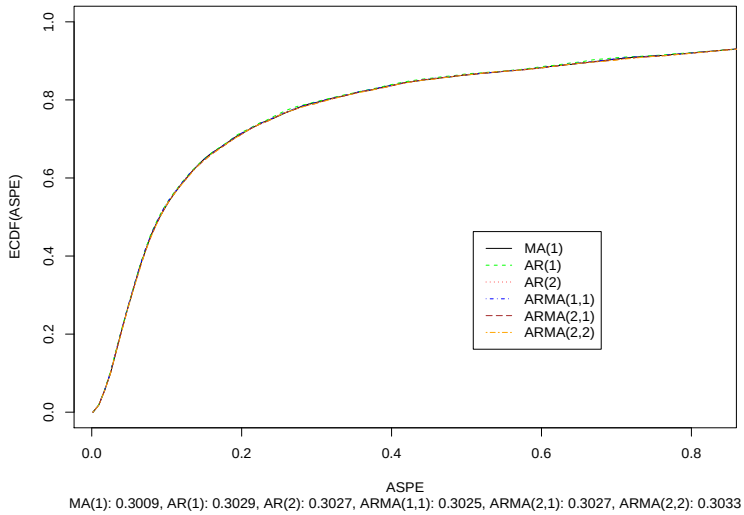

FIGURE 12. Empirical distribution functions of the ASPE for the $S = 10,000$ splits of the first differenced federal funds rate time-series data. Mean values for each model appear in the subtitle below the figure.

to popular time-series tests, our RP test is capable of delivering substantial gains when assessing predictive accuracy. The empirical examples highlight the ease with which the method can be deployed across a range of application domains (cross-section, panel and time-series). We also

present telling empirical evidence as to how overspecified parametric and nonparametric models may not always provide the most accurate approximations to the underlying DGP. Thus, our method can be used as an auxiliary tool for assessing the accuracy of a selected model thereby enhancing any insights one might otherwise glean from empirical exercises.

Fruitful extensions of this approach could include its use in non-regression settings such as the modelling of counts, survival times, or even probabilities. We leave rigorous analysis on optimal selection of the hold-out-sample size and its impact on the resulting test statistic for future research. One could also trivially extend our testing idea to include formal tests of stochastic dominance as opposed to the visual arguments advocated in the paper, though we leave this an an exercise for the interested reader.

References

Anglin, P. M. & Gençay, R. (1996), 'Semiparametric estimation of a hedonic price function', *Journal of Applied Econometrics* **11**, 633–648.

Ashley, R. (2003), 'Statistically significant forecasting improvements: How much out-of-sample data is likely necessary?', *International Journal of Forecasting* **19**, 229–239.

Ashley, R. A. (1998), 'A new technique for postsample model selection and validation', *Journal of Economic Dynamics and Control* **22**, 647–665.

Ashley, R. A., Granger, C. W. J. & Schmalensee, R. (1980), 'Advertising and aggregate consumption: An analysis of causality', *Econometrica* **48**, 1149–1167.

Bauer, D. F. (1972), 'Constructing confidence sets using rank statistics', *Journal of the American Statistical Association* **67**, 687–690.

Bin, O. (2004), 'A prediction comparison of housing sales prices by parametric versus semi-parametric regressions', *Journal of Housing Economics* **13**, 68–84.

Bühlmann, P. (1997), 'Sieve bootstrap for time series', *Bernoulli* **3**, 123–148.

Canty, A. & Ripley, B. (2008), *boot: Bootstrap R (S-Plus) Functions*. R package version 1.2-34.
URL: *http://www.r-project.org*

Chang, Y.-C. & Hung, W.-L. (2007), 'LINEX loss functions with applications to determining the optimum process parameters', *Quality and Quantity* **41**(2), 291–301.

Clark, T. E. & West, K. D. (2007), 'Approximately normal tests for equal predictive accuracy in nested models', *Journal of Econometrics* **138**, 391–311.

Corradi, V. & Swanson, N. R. (2002), 'A consistent test for nonlinear out of sample predictive accuracy', *Journal of Econometrics* **110**, 353–381.

Corradi, V. & Swanson, N. R. (2004), 'Some recent developments in predictive accuracy testing with nested models and (generic) nonlinear alternatives', *International Journal of Forecasting* **20**, 185–199.

Corradi, V. & Swanson, N. R. (2007), 'Nonparametric bootstrap procedures for predictive inference based on recursive estimation schemes', *International Economic Review* **48**(1), 67–109.

Croissant, Y. (2006), *Ecdat: Data sets for econometrics*. R package version 0.1-5.
URL: *http://www.r-project.org*

Davidson, R. & Duclos, J.-Y. (2000), 'Statistical inference for stochastic dominance and for the measurement of poverty and inequality', *Econometrica* **68**, 1435–1464.

Davidson, R. & MacKinnon, J. G. (2002), 'Bootstrap J tests of nonnested linear regression models', *Journal of Econometrics* **109**, 167–193.

Davidson, R. & MacKinnon, J. G. (2004), *Econometric Theory and Methods*, Oxford University Press, New York, NY.

Davison, A. C. & Hinkley, D. V. (1997), *Bootstrap Methods and Their Application*, Cambridge University Press, Cambridge, UK.

De Gooijer, J. G. & Hyndman, R. J. (2006), '25 years of times series forecasting', *International Journal of Forecasting* **22**, 443–473.

Diebold, F. X. & Mariano, R. S. (1995), 'Comparing predictive accuracy', *Journal of Business & Economic Statistics* **13**(3), 253–265.

Durlauf, S. N., Johnson, P. & Temple, J. (2005), Growth econometrics, *in* P. Aghion & S. N. Durlauf, eds, 'Handbook of Economic Growth', Vol. 1A, North-Holland: Amsterdam.

Efron, B. (1982), *The Jackknife, the Bootstrap, and Other Resampling Plans*, Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania 19103.

Efron, B. (1983), 'Estimating the error rate of a prediction rule: Improvement on cross-validation', *Journal of the American Statistical Association* **78**(382), 316–331.

Efron, B. (1986), 'How biased is the apparent error rate of a prediction rule', *Journal of the American Statistical Association* **81**(394), 461–470.

Geisser, S. (1975), 'A predictive sample reuse method with application', *Journal of the American Statistical Association* **70**, 320–328.

Gençay, R. & Yang, X. (1996), 'A prediction comparison of residential housing prices by parametric versus semi-parametric conditional mean estimators', *Economics Letters* **52**, 129–135.

Goodwin, P. (2007), 'Should we be using significance test in forecasting research?', *International Journal of Forecasting* **23**, 333–334.

Granger, C. W. J. & Newbold, P. (1986), *Forecasting economic time series*, Academic Press, San Diego, CA.

Hansen, B. E. (2005), 'Challenges for econometric model selection', *Econometric Theory* **21**, 60–68.

Härdle, W. & Mammen, E. (1993), 'Comparing nonparametric versus parametric regression fits', *Annals of Statistics* **21**(4), 1926–1947.

Härdle, W. & Marron, J. S. (1990), 'Semiparametric comparison of regression curves', *Annals of Statistics* **18**(1), 63–89.

Harvey, D. I., Leybourne, S. J. & Newbold, P. (1997), 'Testing the equality of prediction mean squared errors', *International Journal of Forecasting* **13**, 281–291.

Harvey, D. I., Leybourne, S. J. & Newbold, P. (1998), 'Tests of forecast encompassing', *Journal of Business & Economics Statistics* **16**(2), 254–259.

Harvey, D. I. & Newbold, P. (2000), 'Tests for multiple forecast encompassing', *Journal of Applied Econometrics* **15**, 471–482.

Hayfield, T. & Racine, J. S. (2008), 'Nonparametric econometrics: The np package', *Journal of Statistical Software* **27**(5).
URL: *http://www.jstatsoft.org/v27/i05/*

Henderson, D. J., Papageorgiou, C. & Parmeter, C. F. (2008), Are any growth theories linear? Why we should care about what the evidence tells us. Munich Personal RePEc Archive Paper No. 8767.

Horowitz, J. L. (2004), 'Bootstrap methods for markov processes', *Econometrica* **71**(4), 1049–1082.

Hsiao, C., Li, Q. & Racine, J. S. (2007), 'A consistent model specification test with mixed discrete and continuous data', *Journal of Econometrics* **140**, 802–826.

Hurvich, C. M., Simonoff, J. S. & Tsai, C. L. (1998), 'Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion', *Journal of the Royal Statistical Society Series B* **60**, 271–293.

Hyndman, R. J. (2008), *forecast: Forecasting functions for time series*. R package version 1.19.
URL: *http://www.robjhyndman.com/Rlibrary/forecast/*

Hyndman, R. J. & Koehler, A. B. (2006), 'Another look at measures of forecast accuracy', *International Journal of Forecasting* **22**, 679–688.

Inoue, A. & Kilian, L. (2004), 'In-sample and out-of-sample tests of predictability: Which one should we use?', *Econometric Reviews* **23**, 371–402.

Lahiri, S. N. (2003), *Resampling Methods for Dependent Data*, Springer-Verlag, New York, NY.

Liu, Z. & Stengos, T. (1999), 'Non-linearities in cross country growth regressions: A semiparametric approach', *Journal of Applied Econometrics* **14**, 527–538.

Maasoumi, E., Racine, J. S. & Stengos, T. (2007), 'Growth and convergence: A profile of distribution dynamics and mobility', *Journal of Econometrics* **136**, 483–508.

McCracken, M. W. (2000), 'Robust out-of-sample prediction', *Journal of Econometrics* **99**(2), 195–223.

Meade, N. (2002), 'A comparison of the accuracy of short term foreign exchange forecasting methods', *International Journal of Forecasting* **18**, 67–83.

Medeiros, M. C., Teräsvirta, T. & Rech, G. (2006), 'Building neural network models for time series: a statistical approach', *Journal of Forecasting* **25**, 49–75.

Murphy, K. M. & Welch, F. (1990), 'Empirical age-earnings profiles', *Journal of Labor Economics* **8**(2), 202–229.

Pagan, A. & Ullah, A. (1999), *Nonparametric Econometrics*, Cambridge University Press, New York.

Parmeter, C. F., Henderson, D. J. & Kumbhakar, S. C. (2007), 'Nonparametric estimation of a hedonic price function', *Journal of Applied Econometrics* **22**, 695–699.

Patton, A., Politis, D. N. & White, H. (2008), 'CORRECTION TO "Automatic block-length selection for the dependent bootstrap" by D. Politis and H. White', *Preprint* .

Politis, D. N. & Romano, J. P. (1992), A circular block-resampling procedure for stationary data, *in* R. LePage & R. Billard, eds, 'Exploring the Limits of Bootstrap', John Wiley, New York, pp. 263–270.

Politis, D. N. & Romano, J. P. (1994), 'The stationary bootstrap', *Journal of the American Statistical Association* **89**(428), 1303–1313.

Politis, D. N. & White, H. (2004), 'Automatic block-length selection for the dependent bootstrap', *Econometric Reviews* **23**(1), 53–70.

R Development Core Team (2008), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL: *http://www.R-project.org*

Racine, J. S. (1993), 'An efficient cross–validation algorithm for window width selection for nonparametric kernel regression', *Communications in Statistics* **22**(4), 1107–1114.

Racine, J. S. & Long, L. (2008), A partially linear kernel estimator for categorical data. McMaster University, Department of Economics Working Paper.

Ramsey, J. (1969), 'Tests for specification error in classical linear least squares regression analysis', *Journal of the Royal Statistical Society, Series B* **31**, 350–371.

Shen, X. & Ye, J. (2002), 'Model selection', *Journal of the American Statistical Association* **97**(457), 210–221.

Stone, C. J. (1974), 'Cross-validatory choice and assessment of statistical predictions (with discussion)', *Journal of the Royal Statistical Society* **36**, 111–147.

Stone, C. J. (1984), 'An asymptotically optimal window selection rule for kernel density estimates', *Annals of Statistics* **12**, 1285–1297.

Swanson, N. R. & White, H. (1997), 'A model selection approach to real-time macroeconomic forecasting using linear models and artificial neural networks', *The Review of Economics and Statistics* **79**, 540–550.

van Dijk, D. & Franses, P. H. (2003), 'Selecting a nonlinear time series model using weighted tests of equal forecast accuracy', *Oxford Bulletin of Economics and Statistics* **65**, 727–744.

Wahba, G. & Wold, S. (1975), 'A completely automatic french curve: Fitting spline functions by cross-validation', *Communications in Statistics* **4**, 1–17.

West, K. (1996), 'Asymptotic inference about predictive ability', *Econometrica* **64**, 1067–1084.

West, K. D. & McCracken, M. W. (1998), 'Regression-based tests of predictive ability', *International Economic Review* **39**(3), 817–840.

White, H. (2000), 'A reality check for data snooping', *Econometrica* **68**(5), 1097–1126.

White, H. (2001), *Asymptotic Theory for Econometricians*, Academic Press, San Diego, CA.

Wooldridge, J. M. (1992), 'A test of functional form against nonparametric alternatives', *Econometric Theory* **8**, 452–475.

Wooldridge, J. M. (1994), 'A simple specification test for the predictive ability of transformation models', *The Review of Economics and Statistics* **76**(1), 59–65.

Wooldridge, J. M. (2003), *Introductory Econometrics*, 3 edn, Thompson South-Western, Mason, OH.

Ye, J. (1998), 'On measuring and correcting the effects of data mining and data selection', *Journal of the American Statistical Association* **93**(441), 120–131.