Research in Economics 70 (2016) 304–319

Contents lists available at ScienceDirect

Research in Economics

journal homepage: www.elsevier.com/locate/rie

# A comparison of some out-of-sample tests of predictability in iterated multi-step-ahead forecasts ☆

Pablo M. Pincheira [a], Kenneth D. West [b,*]

[a] School of Business, Adolfo Ibáñez University, Diagonal Las Torres 2640, Peñalolén, Santiago, Chile
[b] Department of Economics, University of Wisconsin-Madison, 1180 Observatory Drive, Madison, WI 53706-1393, USA

ARTICLE INFO

*Article history:*
Received 3 February 2016
Accepted 15 March 2016
Available online 1 April 2016

*Keywords:*
Causality
Random walk
Out-of-sample
Prediction
Multistep forecasts

ABSTRACT

We consider tests of equal population forecasting ability when mean squared prediction error is the metric for forecasting ability, the two competing models are nested, and the iterated method is used to obtain multistep forecasts. We use Monte Carlo simulations to explore the size and power of the MSPE-adjusted test of Clark and West (2006, 2007) (CW) and the Diebold–Mariano–West (DMW) test. The empirical size of the CW test is almost always tolerable: across a set of 252 simulation results that span 5 DGPs, 9 horizons, and various sample sizes, the median size of nominal 10% tests is 8.8%. The comparable figure for the DMW test, which is generally undersized, is 2.2%. An exception for DMW occurs for long horizon forecasts and processes that quickly revert to the mean, in which case CW and DMW perform comparably. We argue that this is to be expected, because at long horizons the two competing models are both forecasting the process to have reverted to its mean. An exception for CW occurs with a nonlinear DGP, in which CW is usually oversized. CW has greater power and greater size adjusted power than does DMW in virtually all DGPs, horizons and sample sizes. For both CW and DMW, power tends to fall with the horizon, reflecting the fact that forecasts from the two competing models both converge towards the mean as the horizon grows. Consistent with these results, in an empirical exercise comparing models for inflation, CW yields many more rejections of equal forecasting ability than does DMW, with most of the rejections occurring at short horizons.


## 1. Introduction

Forecast comparisons have long played a role in evaluation of economic models. A prominent early example was in exchange rate economics. Meese and Rogoff's (1983) demonstration that then-popular exchange rate models forecast no better than a random walk stimulated a huge literature, and made forecast evaluation a common element of evaluation of both reduced form and structural exchange rate models (see Engel et al., 2007). A number of recent examples may be found in volume 2 of the Handbook of Forecasting, whose chapters use forecast comparisons to evaluate, for example, everything from DSGE macro-models (Del Negro and Schorfheide, 2013) to macro-finance models for interest rates (Duffee, 2013) to Phillips curves for inflation (Faust and Wright, 2013).

☆ We thank Michael McCracken for helpful comments and Jose Luis Nolazco for assistance with the data.
* Corresponding author.
*E-mail addresses:* pablo.pincheira@uai.cl (P.M. Pincheira), kdwest@wisc.edu (K.D. West).

http://dx.doi.org/10.1016/j.rie.2016.03.002
1090-9443/

Forecast comparisons potentially involve not only point estimates of measures of forecast quality but also standard errors on cross-model differences in forecast quality. Our paper is concerned with the accuracy of inference about forecast quality once a standard error and $t$-statistic are constructed. In the literature, two leading ways to construct standard errors are the methods proposed in Clark and West (2006, 2007) (hereafter CW) and Diebold and Mariano (1995) and West (1996) (hereafter, DMW).[1] Our aim in this paper is to use Monte Carlo simulations to explore the size and power of the CW and DMW tests at multistep horizons. We are interested in their performance both absolutely and relative to one another.

We use the conventional measure of forecast quality, i.e., mean squared prediction error (MSPE). In our simulations, we calibrate our parameters and sample sizes to macro and financial applications such as weekly or monthly exchange rates or stock returns, quarterly GDP or monthly CPI inflation. Four of our five models are linear, and one is nonlinear. Our artificial data generating processes all involve what are called "nested" models, which compare a simple stripped down null model to an alternative model that adds on regressors whose coefficients are presumed to be zero in the null model. We make multistep forecasts using what is called the "iterated" method. As explained in detail in the next section, this method relies on textbook procedures to make multistep forecasts.

The empirical size of the CW test is almost always tolerable: across a set of 252 simulation results that span 5 DGPs, 9 horizons, and various sample sizes, the median size of nominal 10% tests is 8.8%. The comparable figure for the DMW test, which is generally undersized, is 2.2%. An exception for DMW occurs for long horizon forecasts and processes that quickly revert to the mean, in which case the fact that forecasts from both models have reverted to the mean leads DMW to perform as well as CW. An exception for CW occurs with a nonlinear DGP, in which CW is usually oversized; DMW is also oversized but less so than CW. CW has greater power and greater size adjusted power than does DMW in virtually all DGPs, horizons and sample sizes. Power tends to fall with the horizon, consistent with the fact that both models converge towards forecasting the mean.

Two implications for applied work are to use CW in preference to DMW, and to focus on short horizons, because that is where power is greatest. Indeed, in our empirical exercise comparing two models for inflation, CW yields many more rejections of equal forecasting ability than does DMW, with most of the rejections occurring at short horizons.

The simulation results are broadly similar to those in Clark and West (2006, 2007), who focus on one step ahead rather than multistep predictions. They are also similar to the results in Clark and McCracken (2013b), who also compare multistep forecasts using the iterated method.

Our use of the iterated method to construct long horizon forecasts distinguishes our study from most earlier ones. Most of the research evaluating multistep forecast tests assume predictions are constructed using the "direct" rather than iterated method to forecast (the next section briefly explains the direct method). See for instance, Clark and McCracken (2005a) and Busetti and Marcucci (2013). We distinguish ourselves from the aforementioned Clark and McCracken (2013b) paper via use of different DGPs, horizons and sample sizes. Since, as well, some recent empirical literature (e.g., Faust and Wright (2013) and Pincheira and Gatty (2016)) employs the iterated method for multistep forecasts, there is a need for econometric evaluation of forecast inference techniques when the iterated forecasts are used.

We emphasize that we are testing equal population forecasting ability. That is, the relevant set of applications are ones that use forecast comparisons as a model evaluation technique. This is to be distinguished from tests of equal forecasting ability conditional on a given sample, where one is simply looking for a good forecast. See Clark and McCracken (2013a) for further discussion. While DMW can be used to compare equal population forecasting ability when comparing non-nested models (West, 1996), our application is to nested models. So our simulation results are of questionable relevance to comparisons of non-nested models or comparison of forecasting ability conditional on a given sample.

The rest of the paper is organized as follows. Section 2 outlines CW and DMW and the general econometric environment. Section 3 describes our DGPs and our simulation setup. Section 4 presents simulation evidence showing the size and power performance of the two tests. Section 5 illustrates the use of these tests in an empirical application. Section 6 concludes. An on-line appendix available from the authors contains some additional results omitted from the published paper to save space.

## 2. Econometric setup and forecast evaluation framework

### 2.1. Construction of forecasts

Our linear econometric setup considers nested specifications for a scalar dependent variable $y_{t+1}$ as follows:

$$y_{t+1} = X_t'\beta + e_{1t+1} \quad \text{(model 1: null model)}, \tag{2.1}$$

$$y_{t+1} = X_t'\beta + Z_t'\gamma + e_{2t+1} \quad \text{(model 2: alternative model)}, \tag{2.2}$$

where $e_{1t+1}$ and $e_{2t+1}$ are mean zero and i.i.d.

[1] See West (2006) and Clark and McCracken (2013a) for a discussion of some other methods to construct standard errors.