Tests of Conditional Predictive Ability: Some Simulation Evidence [*]

Michael W. McCracken
Federal Reserve Bank of St. Louis

February 2019

Abstract

In this note we provide simulation evidence on the size and power of tests of predictive ability described in Giacomini and White (2006). Our goals are modest. First, we establish that there exist data generating processes that satisfy the null hypotheses of equal conditional predictive ability. We then consider various parameterizations of these DGPs as a means of evaluating the size and power properties of the proposed tests. While some of our results reinforce those in Giacomini and White (2006), others do not. We recommend against using the fixed scheme when conducting these tests and provide evidence that very large bandwidths are sometimes required when estimating long-run variances.

JEL Nos.: C53, C12, C52
Keywords: prediction, out-of-sample, inference

[*] *McCracken*: Research Division; Federal Reserve Bank of St. Louis; P.O. Box 442; St. Louis, MO 63166; michael.w.mccracken@stls.frb.org. The views expressed herein are solely those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of St. Louis or the Federal Reserve System.

1 Introduction

It has become commonplace to evaluate the quality of a predictive model based upon its ability to forecast in a pseudo-out-of-sample framework. This approach typically consists of splitting the existing time series of data into two portions: an in-sample portion used to estimate model parameters and a second portion used to construct and evaluate the accuracy of forecasts. In the context of point forecasts, the evaluation stage typically consists of (i) constructing forecast errors and (ii) based upon sample averages of functions of these errors, conducting inference related to the quality of the forecasts.

An important step forward in this literature was made by Diebold and Mariano (1995). In the context of comparing the accuracy of two survey-based point forecasts, they show that a t-statistic associated with sample averages of loss differentials from these two models can be asymptotically Normal. Asymptotically valid inference on the mean of the loss differentials is therefore straightforward. This type of statistic was extended by West (1996) to explicitly account for the effect parameter estimation error has on the asymptotic variance of that average loss differential when estimated models are used to form the forecasts. This literature then expanded to a variety of distinct topics including comparisons between nested models (Clark and McCracken 2001, McCracken 2008), cointegrating relationships (Corradi, Olivetti, and Swanson, 2001), dynamic model misspecification (Corradi and Swanson 2007), forecast breakdowns (Giacomini and Rossi, 2009), generated predictors (Goncalves, McCracken, and Perron 2017), generated predictands (Li and Patton, 2018), and many others.

In this literature, when the predictive models use estimated parameters, it is typically the case that the asymptotic theory assumes that the sample sizes used to estimate the model parameters are large enough that the parameter estimates ($\hat{\beta}_t$) are consistent for their population counterparts (β^*). In practice this means test statistics, that are functions of the finite-sample τ-step ahead forecast errors $u_{t+\tau}(\hat{\beta}_t) = \hat{u}_{t+\tau}$, are being used to test null hypotheses related to population-level forecast errors $u_{t+\tau}(\beta^*) = u_{t+\tau}$.

A different approach is taken by Giacomini and White (2006, henceforth GW). They emphasize the need to evaluate the accuracy of forecasts as we observe them in practice which requires admitting that they typically depend on estimated parameters. As an example, when comparing the accuracy of point forecasts from two models ($i = 1, 2$) under quadratic loss, the bulk of the literature focuses on a null hypothesis of the form $E(u^2_{1,t+\tau} - u^2_{2,t+\tau}) = 0$ while GW instead consider hypotheses of the form

$E(\hat{u}^2_{1,t+\tau} - \hat{u}^2_{2,t+\tau}) = 0$ and $E(\hat{u}^2_{1,t+\tau} - \hat{u}^2_{2,t+\tau}|\Im_t) = 0$. The first of these three hypotheses is a statement about the unconditional expectation of the population-level loss differential while the second and third are statements about the unconditional and conditional (based on a time t information set $\Im_t$) expectation of the finite-sample loss differential.[1]

These two new hypotheses required new test statistics and more importantly, a new approach to inference that prevents the parameter estimates from converging to their population counterparts as the total sample size increases. GW achieve this by requiring that a finite number of observations (R) are always used to estimate the parameters.[2] Either the estimates depend on a rolling window of observations x such that $\hat{\beta}_t = \beta(x_{t-R+1}, ..., x_t)$ or a fixed window of observations such that $\hat{\beta}_t = \beta(x_1, ..., x_R)$. With this assumption in hand, and along with other technical assumptions on the loss differential $\hat{d}_{t+\tau} = \hat{u}^2_{1,t+\tau} - \hat{u}^2_{2,t+\tau}$, $t = R, ..., T - \tau = R + P - 1$, they prove that a statistic of the form $P^{-1/2} \sum_{t=R}^{T-\tau} \hat{d}_{t+\tau}/\hat{\omega}$ is asymptotically standard Normal under the null hypothesis of equal finite sample predictive ability (either unconditional $E(\hat{u}^2_{1,t+\tau} - \hat{u}^2_{2,t+\tau}) = 0$ or conditional $E(\hat{u}^2_{1,t+\tau} - \hat{u}^2_{2,t+\tau}|\Im_t) = 0$) so long as a consistent estimate of the long-run variance of $\hat{d}_{t+\tau}$ is used to form $\hat{\omega}$.

In this paper we revisit the results in GW through a small handful of simulations. Our goals are modest. Specifically, we want to investigate the size and power properties of the tests of equal finite sample predictive ability proposed in GW given data generating processes (DGPs) that can be shown to satisfy the null hypotheses. This likely strikes the reader as odd but there exist no simulations in GW that delineate a DGP that satisfies either of the null hypotheses. Instead, in their first simulation, a DGP is developed that satisfies the null hypothesis $E[P^{-1} \sum_{t=R}^{T-\tau} (\hat{u}^2_{1,t+\tau} - \hat{u}^2_{2,t+\tau})] = 0$ and hence *on average* the models are equally accurate. This null hypothesis clearly does not inform us about the null hypothesis $E(\hat{u}^2_{1,t+\tau} - \hat{u}^2_{2,t+\tau}|\Im_t) = 0$ and only loosely informs us about the null hypothesis $E(\hat{u}^2_{1,t+\tau} - \hat{u}^2_{2,t+\tau}) = 0$ for all $t = R, ..., T$.

In a second set of simulations, GW simulate the loss differential directly rather than provide a DGP that implies the loss differentials. That is, they simulate $\hat{d}_{t+1}$ directly as $\hat{d}_{t+1} = \mu(1-\rho) + \rho\hat{d}_t + \varepsilon_{t+1}$, $\varepsilon_{t+1} \sim i.i.d.N(0, 1)$ without providing a DGP that implies that the loss differential forms an $AR(1)$ with drift. These simulations are intended to inform us about the size and power of tests of the null $E(\hat{d}_{t+1}|\Im_t) = 0$. The null is imposed by

[1] See equation (2) on page 1549 of GW.

[2] As a technical matter, they permit R to vary across forecast origins but the maximum is bounded from above. For brevity we assume it is constant.

setting μ and ρ to zero while the alternative is achieved by allowing either to be non-zero. While their simulations indicate the test can be accurately sized and have power against this set of alternatives, the simulation does not provide an example of a DGP in which the null hypothesis can hold. As such, one of our goals is to show that there exists at least one DGP that satisfies the null of equal conditional finite-sample predictive ability when the forecasting exercise requires estimating an unknown parameter.

Our simulation are sometimes, but not always, supportive of the size and power properties reported in GW. When the forecast horizon is one, the nominal size of the test is quite good for tests of conditional, finite-sample predictive ability. But as the horizon grows, size distortions arise due to the need to estimate long-run variances. Size distortions are particularly acute when we consider the tests of unconditional, finite-sample predictive ability. For these tests there are two issues that cause size distortions. The first is that under the fixed scheme, the test is not asymptotically Normal even under the null because the loss differential is not mixing. In fact, the statistic diverges with probability one under the null. The second is that under the rolling scheme, the need to estimate a long-run variance is particularly problematic because even one step ahead forecasts imply loss differentials that are serially correlated of order $R-1$. Rolling window estimation of the parameters induces serial correlation in the loss differentials and properly accounting for that degree of autocorrelation can be difficult in finite samples.

Power comparisons are difficult to make due to the lack of a detailed DGP in GW. Here, we consider two types of alternatives. The first is that the window size used to estimate the parameters ($\tilde{R}$) does not align with that maintained in the DGP (R). For the second, we consider the case in which the fixed (rolling) scheme is used to estimate the model parameters while the DGP is defined using the rolling (fixed) scheme. For the former, reasonable power only arises when the out-of-sample period is quite large and even then it is rarely the case that the actual power is greater than 70%. In many cases, actual power of the test aligns closely with rejection frequencies observed under the null. For the latter alternative, actual power is sometimes improved and sometimes not. Regardless, rejection frequencies greater than 70% are generally restricted to out-of-sample sizes that are quite large.

Before proceeding it is worth noting other contributions to the literature on tests designed to evaluate finite sample predictive ability rather than population predictive ability. Clark and McCracken (2015) develop a test of equal average finite sample predictive ability $E[P^{-1}\sum_{t=R}^{T-\tau}(\hat{u}_{1,t+\tau}^2 - \hat{u}_{2,t+\tau}^2)] = 0$ for both rolling and expanding estimation windows. Simulation evidence is provided using DGPs but the null is distinct from that in GW. Perhaps most importantly, Timmermann and Zhu (2017) develop theoretical results allowing expanding estimation windows when testing the same null hypothesis $E(\hat{u}_{1,t+\tau}^2 - \hat{u}_{2,t+\tau}^2|\Im_t) = 0$ considered in GW. Despite their theoretical results, they do not provide simulations that explicitly delineate DGPs that satisfy the null hypothesis.[3]

The remainder proceeds as follows. Section 2 provides simulation evidence on the size and power of tests of the null hypothesis $E(\hat{u}_{1,t+\tau}^2 - \hat{u}_{2,t+\tau}^2) = 0$ for all $t = R, ..., T - \tau$ based on a fully specified DGP. Section 3 does the same but for the null hypothesis $E(\hat{u}_{1,t+\tau}^2 - \hat{u}_{2,t+\tau}^2|\Im_t) = 0$. Section 4 concludes.

2 Tests of Unconditional, Finite-Sample Predictive Ability

In this section we delineate a DGP that satisfies the null hypothesis $E(\hat{u}_{1,t+\tau}^2 - \hat{u}_{2,t+\tau}^2) = 0$ for all $t = R, ..., T$. We do so under both the fixed and rolling estimation scheme and across a range of forecast horizons τ, sample sizes T, and window sizes R. By varying the parameterization of the DGP we also address the power properties of the test. It's important to note that the null hypothesis requires correct specification of the estimation window $\tilde{R}$. If the estimation window $\tilde{R}$ does not equal the value of R that defines the DGP then the null is not satisfied and constitutes an alternative.

- Our DGP is motivated by simple application in which a no-change forecast is compared to that from a location model with a single estimated parameter. That is, $\hat{y}_{1,t+\tau} = 0$ while $\hat{y}_{2,t+\tau} = \bar{y}_t$ where $\bar{y}_t = \bar{y}_{\tilde{R}}$ for all t under the fixed scheme and $\bar{y}_t = \tilde{R}^{-1}\sum_{s=t-\tilde{R}+1}^{t} y_s$ under the rolling scheme. Straightforward algebra reveals that if $y_t = \mu + \eta_t$, $\eta_t = \varepsilon_t + \sum_{j=1}^{\tau-1}\theta_j\varepsilon_{t-j}$ with $\varepsilon_t \sim i.i.d.N(0, \sigma^2)$ then $E(\hat{u}_{1,t+\tau}^2 - \hat{u}_{2,t+\tau}^2) = 0$ for all $t = R, ..., T$ when $\mu = R^{-1/2}[\gamma_0 + 2\sum_{j=1}^{R-1}\left(\frac{R-j}{R}\right)\gamma_j]^{1/2}$, $\gamma_j = E\eta_t\eta_{t-j}$, and $\tilde{R} = R$.

- In all simulations the test statistic takes the form $P^{-1/2}\sum_{t=\tilde{R}}^{T-\tau}\hat{d}_{t+\tau}/\hat{\omega}$. With this statistic we use standard Normal critical values to conduct a two-sided test of the null. In all tables we only report results associated with a nominal size of 5%.

[3] Timmermann and Zhu (2017) describe a simple DGP involving two non-nested models in the text of their paper but do not consider this example in their simulations. Since the described forecasting exercise does not entail estimating model parameters, the focus of this note, we do not pursue this DGP in our simulations.

Unreported results associated with 1% and 10% provide comparable results and are excluded for brevity.

- We use sample sizes in which R, $\tilde{R}$, and P range across 25, 75, 25, and 175 – similar to those used in GW. We also consider $P = 1000$. In each simulation, the total sample size T is $\tilde{R} + P + \tau - 1$.

- Throughout we set the variance of the primitive shocks σ^2 to 1. Unreported results that set σ^2 to 0.1 and 10 are quite similar and excluded for brevity. The MA coefficients are set to $\theta_j = (0.5)^j$ in all simulations.

- We consider forecast horizons τ of 1, 3, and 12. Unreported results that set τ to 24 are comparable to those for 12 and excluded for brevity.

- In all results we use the Bartlett kernel to estimate the long-run variance (i.e. Newey and West, 1987). The choice of bandwidth does have an affect and hence we experimented with a few including fixed values of τ and $\tilde{R}$ but also a sample dependent version $\lfloor 4(P/100)^{2/9} \rfloor + 1$ (Newey and West, 1994; Andrews and Monihan, 1992). With $\tau - 1$ lags, the actual size was better (than using $\tilde{R} - 1$) for smaller samples because it entailed estimating fewer autocovariances but was poor for the larger samples because it didn't estimate enough of them. With $\tilde{R} - 1$ lags, it was the reverse: actual size was worse (than using $\tau - 1$) in smaller samples but improved as the sample got larger because it was estimating the correct numbers of autocovariances.[4] Broadly speaking, all size results were (weakly) more accurate using the sample dependent version and hence all reported results use that bandwidth.

- In all experiments the number of replications is $5,000$.

2.1 Fixed

We begin by investigating size and power of the test when the fixed scheme is used. In Table 1, rows that are in bold are those for which the null hypothesis holds. All other values represent alternative hypotheses driven by the fact that $R \neq \tilde{R}$. In all cases we find that the actual size of the test is much higher than the nominal 5% level. Making it worse is the fact that the rejection frequencies generally get worse, not better, as the sample size increases.

[4] Note too that when $\tilde{R} > P$, a bandwidth of $\tilde{R}$ observations is infeasible because there are no observations available to estimate, for example, the $\tilde{R} - 1$th autocovariance.

Given our simple DGPs, a bit of algebra allows us to identify the root of the size distortions. For simplicity let $\tau = 1$, $R = \tilde{R}$, and $\hat{\omega}^2 = P^{-1}\sum_{t=R}^{T-1}\hat{d}_{t+1}^2$. Rearranging terms we find that $P^{-1/2}\sum_{t=R}^{T-1}\hat{d}_{t+1}/\hat{\omega}$ equals

$$\frac{2(P^{-1/2}\sum_{t=R}^{T-1}\varepsilon_{t+1})\bar{y}_R + P^{1/2}(\mu^2 - (R^{-1}\sum_{s=1}^{R}\varepsilon_s)^2)}{\sqrt{4\bar{y}_R^2(P^{-1}\sum_{t=R}^{T-1}\varepsilon_{t+1}^2) + 4\bar{y}_R^2(2\mu - \bar{y}_R)(P^{-1}\sum_{t=R}^{T-1}\varepsilon_{t+1}) + (\mu^2 - (R^{-1}\sum_{s=1}^{R}\varepsilon_s)^2)^2}}.$$

As P diverges, holding R constant, the denominator converges to a finite, albeit stochastic, constant $\sqrt{4\bar{y}_R^2\sigma^2 + (\mu^2 - (R^{-1}\sum_{s=1}^{R}\varepsilon_s)^2)^2}$. In contrast, while the first part of the numerator is asymptotically Normal, the second part diverges to $\pm\infty$ with probability one. The fundamental issue that makes the test inconsistent is that the loss differential isn't mixing. While it is true that under the fixed scheme, the loss differential $\hat{d}_{t+1}$ is a finite dimensioned function of mixing variables (i.e. $y_{t+1}, y_R, ..., y_1$), the window size $(t+1) - 1$ is increasing and hence Lemma 2.1 of White and Domowitz (1984) doesn't apply.[5]

Since the actual size of the test is so poor, it is difficult to interpret how powerful the test is when $R \neq \tilde{R}$. For the samples in which $P < 1000$, the rejection frequencies are often similar to those under the null unless $\tilde{R} = 25$ but even there actual power tops off around 30%. The rejection frequencies under the alternative rise to 60% and more when $P = 1000$ but the size distortions are increasing as well making interpretation difficult.

2.2 Rolling

In Table 2 we report similar results but when the rolling scheme is used. Here we find that the actual size of the test is much more in line with what we would expect from a nominally 5% test. At the shortest forecast horizon, $\tau = 1$, rejection frequencies are as high as 9% when $P = 25$ but otherwise range between 3% and 7%. At the longer horizons, the actual size of the test is as high as 20% but improve substantially as P increases.

The size distortions are due in part to the need to estimate a long-run variance that accounts for serial correlation in the loss differential. This serial correlation takes the form of an $MA(\tilde{R} - 1)$. To see this, consider the numerator of the test statistic when $R = \tilde{R}$, and $\tau = 1$. Rearranging terms we find that $P^{-1/2}\sum_{t=R}^{T-1}\hat{d}_{t+1}$ equals

$$2(P^{-1/2}\sum_{t=R}^{T-1}\varepsilon_{t+1}\bar{y}_t) + P^{-1/2}\sum_{t=R}^{T-1}(\mu^2 - (R^{-1}\sum_{s=t-R+1}^{t}\varepsilon_s)^2).$$

[5] In the notation of White and Domowitz (1984), their tuning parameter τ is equal to $t-1$ in this simulation. Since they assume that τ is finite, and in our exercise t is arbitrarily large, Lemma 2.1 does not apply.

Both right hand side terms are zero mean and mixing and hence the numerator is asymptotically Normal. However, a closer look reveals that $\hat{d}_{t+1}$ is much more persistent than one might have expected. The rolling windows approach induces serial correlation of order $\tilde{R}-1$ which is substantially more persistent than one might have guessed given that the forecasts are at the one-step horizon. This, in turn, entails estimating a large number of autocovariances when forming a consistent estimate of the long-run variance.

Under the alternative in which $\tilde{R} \neq R$, power is quite poor with rejection frequencies generally below 20% for all but the largest sample size. In a few cases the test is biased: rejection frequencies under the alternative are lower than under the null. For the largest sample size, actual power can be as large as 80% but remains as low as 10% for some permutations of $\tilde{R} \neq R$.

3 Tests of Conditional, Finite-Sample Predictive Ability

In this section we delineate a DGP that satisfies the null hypothesis $E(\hat{u}^2_{1,t+\tau} - \hat{u}^2_{2,t+\tau}|\Im_t) = 0$. In contrast to the previous section, the DGP is specific to whether the fixed or rolling scheme is being used. This means that if the DGP is designed relative to the fixed (rolling) scheme, but a rolling (fixed) model is used for forecasting, there must be a deviation from the null hypothesis. In addition, since $R \neq \tilde{R}$ continues to indicate a deviation from the null hypothesis, we provide simulations on power of the test against these two forms of alternatives.

Our DGPs are again developed based on a simple application in which a no-change forecast is compared to that from a location model with a single estimated parameter. That is, $\hat{y}_{1,t+\tau} = 0$ while $\hat{y}_{2,t+\tau} = \bar{y}_t$ where $\bar{y}_t = \bar{y}_{\tilde{R}}$ for all t under the fixed scheme and $\bar{y}_t = \tilde{R}^{-1}\sum_{s=t-\tilde{R}+1}^{t} y_s$ under the rolling scheme.

- Fixed scheme. For $t = 1, ..., R$ set $y_t = 2\eta_t$ where $\eta_t = \varepsilon_t + \sum_{j=1}^{\tau-1}\theta_j\varepsilon_{t-j}$ with $\varepsilon_t \sim i.i.d.N(0, \sigma^2)$. For $t = R+1, ..., T$ set $y_t = \frac{1}{2}\bar{y}_R + \eta_t$.

- Rolling scheme. For $t = 1, ..., T$ generate y_t as a stationary $ARMA(R+h-1, h-1)$ with moving average parameters θ_j, $j = 1, ..., h-1$, and autoregressive parameters set so that the first $h-1$ values are zero (i.e. $\alpha_j = 0$ for $j = 1, ..., h-1$) and the remaining R values are $\frac{1}{2R}$ (i.e. $\alpha_j = \frac{1}{2R}$ for $j = h, ..., R+h-1$).[6]

[6]The intitial conditions are set to zero and a burn-in of $10,000$ is used.

- As we did for the unconditional tests, we set $\sigma^2 = 1$, $\theta_j = (0.5)^j$, let $\tau = 1, 3, 12$, and use the same values of R, $\tilde{R}$, and P.

- We consider two test statistics. The first is simply the baseline t-statistic $P^{-1/2} \sum_{t=\tilde{R}}^{T-\tau} \hat{d}_{t+\tau}/\hat{\omega}$ which is asymptotically standard Normal under the null. But as noted in GW, since the null hypothesis implies $\hat{d}_{t+\tau}$ is uncorrelated with any observable $z_t \in \Im_t$, $P^{-1/2} \sum_{t=\tilde{R}}^{T-\tau} \hat{d}_{t+\tau} z_t$ will also be asymptotically Normal under the null hypothesis. Since z_t is allowed to be non-scalar they suggest a Wald-statistic of the form $(P^{-1/2} \sum_{t=\tilde{R}}^{T-\tau} \hat{d}_{t+\tau} z_t)' \hat{\Omega}^{-1} (P^{-1/2} \sum_{t=\tilde{R}}^{T-\tau} \hat{d}_{t+\tau} z_t)$ which will be asymptotically chi-square with $\dim(z_t)$ degrees of freedom under the null so long as $\hat{\Omega}$ is consistent for the long-run variance of $\hat{d}_{t+\tau} z_t$. Again following the suggestion in GW, we investigate the usefulness of this statistic using $z_t = (1, \hat{d}_t)'$ as the instrument vector.

- When constructing $\hat{\Omega}$, we follow a slightly different approach than when we did for the unconditional tests. When $\tau = 1$, there is no serial correlation in either $\hat{d}_{t+1}$ or $\hat{d}_{t+1} z_t$ and hence their variance is the long-run variance. We therefore construct $\hat{\Omega}$ as $P^{-1} \sum_{t=\tilde{R}}^{T-1} \hat{d}_{t+1}^2 z_t z_t'$ and $\hat{\omega}^2$ as $P^{-1} \sum_{t=\tilde{R}}^{T-1} \hat{d}_{t+1}^2$ when $\tau = 1$. For the longer horizons, we again used the Bartlett kernel and experimented with a bandwidth of τ and the data-dependent version. Since the actual size of the tests were weakly better using the data dependent rule we use that bandwidth when $\tau > 1$.

3.1 Fixed

In Table 3 we provide results associated with the scalar version of the test when the fixed scheme is used in the DGP and in the modeling. The layout of the table is the same as for the previous tables and hence rows in bold are actual size while those not in bold are actual power. In contrast to the results for the unconditional test, the fixed scheme can exhibit reasonable size in large enough samples. When $\tau = 1$, actual size is quite good for all sample sizes. But as we increase the forecast horizon, the need to estimate long-run variances again induces size distortions. Even so, it is comforting to see that for $\tau = 3$, the distortions diminish quickly as the sample size increases. At the longest horizon $\tau = 12$, the actual size improves as the sample increases but remains elevated even when $P = 1000$.

Under the alternative in which $\tilde{R} \neq R$ and the fixed model is used to form forecasts (Table 3, not bold) power is quite poor with rejection frequencies generally below 20% for all but the largest sample size. For the largest sample size, actual power can be as large as 75% but remains as low as 10% for some permutations of $\tilde{R} \neq R$. Under the alternative in which the rolling model is used to form forecasts (Table 4), actual power is typically the same as or even worse than in Table 3 holding $(\tilde{R}, P)$ constant.

In Tables 5 and 6 we repeat the same experiments as in Tables 3 and 4 but now using the Wald form of the test with $z_t = (1, \hat{d}_t)'$ as the instrument vector. In Table 5, the actual size of the test is not wholly unreasonable when $\tau = 1$ though there is a significant amount of under-rejection when $R > 25$. At the longer horizons, actual size of the test is typically poor especially at the smaller sample sizes with rejection frequencies as high as 50%. Even so, consistent with the theory, actual size generally improves as the sample size increases.

Under the alternative in which $\tilde{R} \neq R$, and the fixed model is used to form forecasts (Table 5, not bold), actual power is poor when $\tau = 1$ with rejection frequencies generally below 15% unless $\tilde{R} = 25$ or $P = 1000$ and $\tilde{R} < R$ in which case actual power can be as high as 65%. When $\tau > 1$, actual power is improved at all sample sizes particularly when $\tilde{R} < R$ with rejections frequencies as high as 75%. Across all horizons there is a tendency for actual power to be higher for smaller values of $\tilde{R}$.

Under the alternative in which the rolling model is used to form forecasts (Table 6), actual power is a mixed bag. When $\tau = 1$, actual power is only above 20% when $\tilde{R} = 25 < R$ or when $P = 1000$ and $R = \tilde{R} = 25$. In these cases the rejection frequencies typically lie between 30% and 50% but get as high as 70%. Actual power is uniformly higher at the longer horizons holding $(\tilde{R}, P)$ constant. As was the case for $\tau = 1$, actual power tends to rise as $\tilde{R}$ becomes smaller with rejection frequencies as high as 88%. Somewhat oddly, there are times when the rejection frequencies are u-shaped as P increases for a fixed $\tilde{R}$. As an example, for $\tau = 3$, when $R = 25$ and $\tilde{R} = 75$ the rejection frequencies range from 29%, 20%, 17%, 17%, to 36% as P ranges from 25, 75, 125, 175, to 1000.

3.2 Rolling

In this section, the simulations parallel those in the previous subsection but when the rolling scheme is used to define the DGPs. In Table 7, we report rejection frequencies when both the DGP and model are based on the rolling scheme and the scalar test statistic is used for inference. Under the null, actual size of the test is quite good when $\tau = 1$. As the horizon increases, size distortions arise, especially when P is small but again, in alignment with the theory, actual size improves substantially as the sample size P increases. Even so, with samples as large as $P = 1000$, rejection frequencies remain elevated especially for smaller R.

As it was for the fixed case in Table 3, under the alternative in which $\tilde{R} \neq R$ and the rolling model is used to form forecasts (Table 7, not bold) power is often poor with rejection frequencies typically below 20% for all but the largest sample size. For the largest sample size, actual power can be as large as 80% but remains as low as 5% for some permutations of $\tilde{R} \neq R$. Under the alternative in which the rolling model is used to form forecasts (Table 8), actual power is typically a bit better than in Table 7 holding $(\tilde{R}, P)$ constant.

In Tables 9 and 10 we repeat the same experiments as in Tables 7 and 8 but now using the Wald form of the test with $z_t = (1, \hat{d}_t)'$ as the instrument vector. In Table 9, the actual size is modestly to severely undersized as it was in Table 5 for the fixed scheme. At the longer horizons, actual size of the test is typically poor at the smaller sample sizes with rejection frequencies as high as 40%. Even so, consistent with the theory, actual size generally improves as the sample size increases.

Under the alternative in which the fixed model is used to form forecasts (Table 10), actual power again has a wide range. When $\tau = 1$, power only once reaches 30% for $P < 1000$ and is often a low as 3%. But as the sample size P rises to 1000, actual power improves to as much as 60%. As the horizon increases, power increases almost uniformly holding $(\tilde{R}, P)$ constant. For all horizons, actual power typically improves as $\tilde{R}$ declines. And as was the case for the fixed DGP, there are odd instances in which the rejection frequencies are u-shaped for a fixed value of $\tilde{R}$ as P increases.

4 Conclusion

In this paper we have modest goals. The first goal is simply to provide an example of DGPs that can be shown to satisfy the null hypotheses delineated in GW. This may seem trivial but there exists no such simulation evidence in the literature. In no small part this has arisen because the null hypothesis of equal conditional finite sample predictive ability is extremely narrow and imposes very strong restrictions on the DGP. As shown in GW, given two sequences of point forecasts $\hat{y}_{1,t+\tau}$ and $\hat{y}_{2,t+\tau}$ and assuming a quadratic loss function, the DGP for y must satisfy $y_{t+\tau} = \frac{1}{2}(\hat{y}_{1,t+\tau} + \hat{y}_{2,t+\tau}) + \eta_{t+\tau}$ for $\eta_{t+\tau} \sim MA(\tau - 1)$. We have been able to delineate such a DGP but only for an application in which a no-change forecast is being compared to one formed using a location model. We have not been able to establish an example that permits a broader collection of regression models.

Nevertheless, given our DGP, our second goal is to provide simulation based evidence on the finite sample size and power of the tests suggested in GW. Our results only sometimes reinforce the conclusions provided in GW. The tests can be accurately sized given large enough samples, especially at shorter horizons. They too can exhibit substantial power when deviations from the null are large or the sample size is large. But by providing a DGP-based set of simulations we are able to diagnose details that are not particularly obvious in those provided by GW. The most important of which is simply that the proposed test statistic $P^{-1/2}\sum_{t=\tilde{R}}^{T-\tau}\hat{d}_{t+\tau}/\hat{\omega}$, is not asymptotically valid under the fixed scheme when testing the null of equal unconditional finite-sample predictability. In our simulations, the test statistic diverges under the null and there is reason to believe that this will hold more broadly since the loss differential is not mixing. We also find that for this null, by construction, the rolling scheme induces very high orders of serial correlation in the loss differentials. As such, a large bandwidth is required when estimating the long-run variance which, unfortunately, introduces estimation error and causes size distortions.

References

Andrews, Donald W. K., and J. Christopher Monahan (1992), "An Improved Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimator," *Econometrica* 60, 953–966.

Chao, John C., Corradi, Valentina, and Norman R. Swanson (2001), "An Out of Sample Test of Granger Causality" *Macroeconomic Dynamics* 5, 598-620.

Clark, Todd E., and Michael W. McCracken (2001), "Tests of Equal Forecast Accuracy and Encompassing for Nested Models," *Journal of Econometrics* 105, 85-110.

Clark, Todd E., and Michael W. McCracken (2015), "Nested Forecast Model Comparisons: A New Approach to Testing Equal Accuracy," *Journal of Econometrics* 186, 160-177.

Corradi, Valentina, Olivetti, Claudia, and Norman R. Swanson (2001), "Predictive Ability With Cointegrated Variables," *Journal of Econometrics* 104, 315-358.

Diebold, Francis X., and Roberto S. Mariano (1995), "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics* 13, 253-263.

Giacomini, Rafaella, and Barbara Rossi (2009), "Detecting and Predicting Forecast Breakdowns," *Review of Economic Studies* 76, 669-705.

Giacomini, Rafaella, and Halbert White (2006), "Tests of Conditional Predictive Ability," *Econometrica* 74, 1545-1578.

Goncalves, Silvia, Perron, Benoit, and Michael W. McCracken (2015), "Tests of Equal Accuracy for Nested Models with Estimated Factors," *Journal of Econometrics* 198, 231-252.

Li, Jia, and Andrew Patton (2018), "Asymptotic Inference about Predictive Accuracy Using High Frequency Data," *Journal of Econometrics* 203, 223-240.

McCracken, Michael W. (2007), "Asymptotics for Out-of-Sample Tests of Granger Causality," *Journal of Econometrics* 140, 719-752.

Newey, Whitney K., and Kenneth D. West (1987). "A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica* 55, 703–708.

Newey, Whitney K., and Kenneth D. West (1994), "Automatic Lag Selection in Covariance Matrix Estimation," *Review of Economic Studies* 61, 631-653.

Timmermann, Allan and Yinchu Zhu (2017), "Tests of Forecasting Performance and Choice of Estimation Window," manuscript.

West, Kenneth D. (1996), "Asymptotic Inference About Predictive Ability," *Econometrica* 64, 1067-1084.

White, Halbert (2000), "A Reality Check For Data Snooping," *Econometrica* 68, 1097-1127.

White, Halbert and Ian Domowitz (1984), "Nonlinear Regression with Dependent Observations," *Econometrica* 52, 142-162.

Tables

Table 1: Unconditional: Fixed Model

		$\tau = 1$					$\tau = 3$					$\tau = 12$				
		P					P					P				
R	$\tilde{R}$	25	75	125	175	1000	25	75	125	175	1000	25	75	125	175	1000
25	**25**	**0.15**	**0.24**	**0.33**	**0.39**	**0.70**	**0.23**	**0.29**	**0.37**	**0.41**	**0.72**	**0.28**	**0.33**	**0.40**	**0.46**	**0.73**
	75	0.13	0.19	0.26	0.32	0.74	0.20	0.26	0.31	0.37	0.74	0.25	0.28	0.34	0.40	0.75
	125	0.13	0.19	0.25	0.30	0.76	0.20	0.23	0.28	0.35	0.77	0.26	0.28	0.32	0.40	0.79
	175	0.13	0.17	0.24	0.30	0.78	0.19	0.22	0.29	0.34	0.80	0.26	0.27	0.32	0.39	0.80
75	25	0.13	0.18	0.25	0.28	0.62	0.20	0.23	0.29	0.34	0.63	0.24	0.28	0.32	0.36	0.66
	75	**0.11**	**0.13**	**0.17**	**0.19**	**0.51**	**0.17**	**0.17**	**0.20**	**0.24**	**0.54**	**0.21**	**0.20**	**0.23**	**0.27**	**0.55**
	125	0.10	0.11	0.14	0.17	0.48	0.17	0.15	0.19	0.22	0.52	0.21	0.20	0.21	0.25	0.53
	175	0.10	0.12	0.14	0.17	0.47	0.17	0.16	0.17	0.21	0.49	0.22	0.20	0.20	0.23	0.52
125	25	0.12	0.17	0.23	0.27	0.59	0.19	0.22	0.27	0.31	0.60	0.25	0.27	0.31	0.35	0.64
	75	0.11	0.11	0.14	0.18	0.44	0.18	0.16	0.19	0.21	0.48	0.22	0.20	0.21	0.23	0.50
	125	**0.10**	**0.10**	**0.13**	**0.14**	**0.39**	**0.16**	**0.15**	**0.16**	**0.17**	**0.42**	**0.20**	**0.18**	**0.19**	**0.20**	**0.46**
	175	0.10	0.10	0.12	0.13	0.38	0.17	0.14	0.15	0.16	0.40	0.21	0.18	0.18	0.20	0.43
175	25	0.12	0.17	0.22	0.27	0.57	0.20	0.22	0.26	0.31	0.59	0.25	0.25	0.28	0.33	0.62
	75	0.11	0.11	0.13	0.15	0.43	0.16	0.15	0.18	0.20	0.45	0.21	0.18	0.21	0.22	0.47
	125	0.10	0.10	0.12	0.13	0.37	0.17	0.14	0.15	0.17	0.39	0.21	0.18	0.17	0.20	0.41
	175	**0.09**	**0.10**	**0.11**	**0.12**	**0.34**	**0.15**	**0.13**	**0.15**	**0.15**	**0.36**	**0.22**	**0.17**	**0.16**	**0.18**	**0.40**

Notes: R is the window size used to generate the time series, $\tilde{R}$ is the window size used to estimate the model parameters before generating a forecast, τ is the forecast horizon, and P is the number of forecast periods. Each entry in the table represents the fraction of replications where the null hypothesis was rejected at the 5% level using the standard Normal critical values of a two-sided test for a given R, $\tilde{R}$, τ, and P. Bolded rows indicate when the null hypothesis holds (i.e. when $R = \tilde{R}$). The test statistic takes the form shown in the second bullet of Section 2 (page 4). The long-run variance used in the calculation of the test statistic was estimated using the Bartlett kernel with the bandwidth set to $\lfloor 4(P/100)^{2/9} \rfloor + 1$.

Table 2: Unconditional: Rolling Model

		$\tau = 1$					$\tau = 3$					$\tau = 12$				
		P					P					P				
R	$\tilde{R}$	25	75	125	175	1000	25	75	125	175	1000	25	75	125	175	1000
25	**25**	**0.07**	**0.05**	**0.05**	**0.04**	**0.03**	**0.12**	**0.07**	**0.06**	**0.05**	**0.04**	**0.20**	**0.11**	**0.10**	**0.09**	**0.07**
	75	0.08	0.07	0.06	0.05	0.41	0.16	0.11	0.09	0.10	0.45	0.22	0.16	0.14	0.15	0.48
	125	0.10	0.09	0.10	0.12	0.67	0.17	0.14	0.15	0.16	0.71	0.24	0.20	0.20	0.20	0.73
	175	0.11	0.12	0.14	0.15	0.80	0.18	0.16	0.17	0.20	0.81	0.24	0.22	0.23	0.24	0.82
75	25	0.09	0.11	0.12	0.15	0.51	0.12	0.12	0.13	0.17	0.50	0.21	0.16	0.16	0.18	0.47
	75	**0.08**	**0.06**	**0.06**	**0.05**	**0.03**	**0.13**	**0.09**	**0.07**	**0.06**	**0.03**	**0.19**	**0.12**	**0.10**	**0.08**	**0.06**
	125	0.08	0.06	0.05	0.04	0.03	0.13	0.09	0.08	0.06	0.05	0.19	0.13	0.10	0.09	0.07
	175	0.09	0.06	0.06	0.05	0.06	0.15	0.10	0.08	0.08	0.08	0.20	0.13	0.12	0.10	0.11
125	25	0.10	0.13	0.17	0.20	0.72	0.12	0.14	0.17	0.20	0.70	0.19	0.18	0.18	0.23	0.67
	75	0.09	0.08	0.06	0.07	0.08	0.13	0.09	0.08	0.08	0.10	0.19	0.12	0.11	0.10	0.12
	125	**0.08**	**0.07**	**0.05**	**0.05**	**0.03**	**0.12**	**0.09**	**0.07**	**0.07**	**0.04**	**0.18**	**0.13**	**0.10**	**0.08**	**0.05**
	175	0.09	0.07	0.06	0.05	0.02	0.13	0.10	0.08	0.07	0.03	0.19	0.13	0.10	0.11	0.05
175	25	0.10	0.14	0.20	0.23	0.82	0.13	0.14	0.20	0.23	0.80	0.20	0.18	0.22	0.26	0.74
	75	0.09	0.08	0.08	0.08	0.15	0.13	0.10	0.09	0.10	0.16	0.19	0.13	0.10	0.11	0.19
	125	0.08	0.07	0.07	0.06	0.05	0.14	0.09	0.08	0.07	0.06	0.19	0.12	0.10	0.09	0.08
	175	**0.09**	**0.07**	**0.06**	**0.06**	**0.03**	**0.13**	**0.10**	**0.08**	**0.07**	**0.04**	**0.19**	**0.12**	**0.10**	**0.09**	**0.05**

Notes: See notes to Table 1.

Table 3: Conditional: Fixed DGP and Fixed Model, Scalar Form

		$\tau = 1$					$\tau = 3$					$\tau = 12$				
		P					P					P				
R	$\tilde{R}$	25	75	125	175	1000	25	75	125	175	1000	25	75	125	175	1000
25	**25**	**0.06**	**0.06**	**0.06**	**0.05**	**0.05**	**0.14**	**0.10**	**0.08**	**0.08**	**0.07**	**0.19**	**0.13**	**0.11**	**0.10**	**0.08**
	75	0.09	0.11	0.15	0.18	0.48	0.16	0.17	0.20	0.22	0.51	0.23	0.21	0.22	0.25	0.52
	125	0.08	0.12	0.17	0.21	0.51	0.17	0.19	0.23	0.25	0.54	0.23	0.22	0.26	0.28	0.55
	175	0.09	0.13	0.18	0.22	0.53	0.18	0.19	0.22	0.26	0.55	0.23	0.23	0.25	0.29	0.57
75	25	0.10	0.18	0.28	0.37	0.70	0.18	0.25	0.35	0.41	0.73	0.23	0.32	0.39	0.46	0.73
	75	**0.06**	**0.05**	**0.05**	**0.05**	**0.05**	**0.15**	**0.10**	**0.08**	**0.08**	**0.06**	**0.19**	**0.13**	**0.11**	**0.10**	**0.08**
	125	0.06	0.07	0.07	0.07	0.20	0.15	0.12	0.12	0.11	0.24	0.19	0.14	0.14	0.13	0.25
	175	0.07	0.07	0.08	0.09	0.25	0.16	0.12	0.12	0.12	0.28	0.20	0.16	0.14	0.15	0.32
125	25	0.09	0.14	0.22	0.31	0.71	0.18	0.20	0.27	0.36	0.73	0.25	0.24	0.33	0.40	0.75
	75	0.07	0.10	0.15	0.19	0.45	0.16	0.17	0.21	0.23	0.48	0.21	0.22	0.25	0.27	0.50
	125	**0.07**	**0.06**	**0.05**	**0.05**	**0.05**	**0.15**	**0.11**	**0.09**	**0.09**	**0.06**	**0.19**	**0.14**	**0.11**	**0.10**	**0.08**
	175	0.07	0.06	0.06	0.06	0.12	0.15	0.10	0.09	0.10	0.14	0.20	0.13	0.12	0.13	0.16
175	25	0.09	0.13	0.20	0.26	0.70	0.18	0.20	0.25	0.32	0.71	0.24	0.24	0.27	0.35	0.73
	75	0.07	0.09	0.11	0.16	0.51	0.17	0.14	0.17	0.22	0.53	0.21	0.17	0.20	0.25	0.55
	125	0.07	0.08	0.11	0.14	0.32	0.16	0.14	0.15	0.19	0.35	0.19	0.18	0.18	0.20	0.37
	175	**0.06**	**0.05**	**0.05**	**0.05**	**0.05**	**0.14**	**0.09**	**0.09**	**0.08**	**0.07**	**0.19**	**0.13**	**0.12**	**0.11**	**0.09**

Notes: See notes to Table 1. For $\tau = 1$, the bandwidth of the Bartlett kernel used to estimate the long-run variance was set to 1. Otherwise, the same sample-dependent bandwidth of $\lfloor 4(P/100)^{2/9} \rfloor + 1$ was used.

Table 4: Conditional: Fixed DGP and Rolling Model, Scalar Form

		$\tau = 1$					$\tau = 3$					$\tau = 12$				
		P					P					P				
R	$\tilde{R}$	25	75	125	175	1000	25	75	125	175	1000	25	75	125	175	1000
25	25	0.06	0.08	0.11	0.14	0.54	0.13	0.13	0.16	0.20	0.57	0.20	0.17	0.19	0.23	0.56
	75	0.07	0.08	0.11	0.14	0.42	0.15	0.13	0.17	0.19	0.44	0.21	0.18	0.18	0.22	0.46
	125	0.07	0.11	0.13	0.16	0.42	0.16	0.16	0.18	0.20	0.44	0.22	0.21	0.21	0.23	0.46
	175	0.08	0.11	0.15	0.18	0.44	0.17	0.18	0.19	0.22	0.47	0.21	0.21	0.24	0.25	0.48
75	25	0.07	0.12	0.15	0.19	0.63	0.13	0.18	0.20	0.25	0.67	0.20	0.21	0.22	0.27	0.66
	75	0.06	0.05	0.06	0.06	0.25	0.13	0.10	0.10	0.10	0.30	0.19	0.13	0.12	0.12	0.32
	125	0.05	0.06	0.06	0.07	0.22	0.13	0.10	0.10	0.10	0.25	0.19	0.14	0.12	0.12	0.28
	175	0.06	0.07	0.06	0.07	0.22	0.14	0.10	0.10	0.11	0.24	0.19	0.15	0.13	0.13	0.26
125	25	0.07	0.12	0.17	0.23	0.69	0.14	0.17	0.24	0.28	0.71	0.20	0.20	0.26	0.30	0.70
	75	0.06	0.09	0.09	0.10	0.29	0.13	0.11	0.13	0.14	0.32	0.18	0.15	0.16	0.16	0.35
	125	0.06	0.06	0.05	0.06	0.16	0.13	0.09	0.09	0.09	0.21	0.17	0.12	0.11	0.11	0.23
	175	0.06	0.05	0.06	0.05	0.15	0.13	0.09	0.09	0.09	0.17	0.18	0.13	0.11	0.11	0.20
175	25	0.07	0.12	0.18	0.25	0.72	0.14	0.18	0.25	0.32	0.75	0.20	0.20	0.25	0.33	0.74
	75	0.06	0.07	0.09	0.11	0.30	0.13	0.11	0.13	0.14	0.35	0.18	0.14	0.15	0.18	0.36
	125	0.06	0.07	0.08	0.08	0.18	0.13	0.11	0.10	0.11	0.23	0.18	0.13	0.12	0.14	0.24
	175	0.06	0.05	0.05	0.06	0.13	0.13	0.08	0.09	0.08	0.15	0.18	0.12	0.10	0.11	0.17

Notes: See notes to Table 1. For $\tau = 1$, the bandwidth of the Bartlett kernel used to estimate the long-run variance was set to 1. Otherwise, the same sample-dependent bandwidth of $\lfloor 4(P/100)^{2/9} \rfloor + 1$ was used. No values are bolded because the null hypothesis does not hold for any entry.

Table 5: Conditional: Fixed DGP and Fixed Model, Wald Form

		$\tau = 1$					$\tau = 3$					$\tau = 12$				
				P					P					P		
R	$\tilde{R}$	25	75	125	175	1000	25	75	125	175	1000	25	75	125	175	1000
25	**25**	**0.08**	**0.07**	**0.06**	**0.07**	**0.07**	**0.39**	**0.31**	**0.30**	**0.30**	**0.27**	**0.54**	**0.38**	**0.37**	**0.35**	**0.34**
	75	0.06	0.07	0.09	0.11	0.38	0.31	0.25	0.26	0.27	0.47	0.46	0.33	0.31	0.32	0.51
	125	0.06	0.07	0.11	0.12	0.41	0.28	0.22	0.25	0.27	0.47	0.45	0.30	0.30	0.31	0.52
	175	0.05	0.07	0.11	0.14	0.43	0.26	0.23	0.23	0.25	0.49	0.43	0.29	0.30	0.30	0.50
75	25	0.27	0.29	0.32	0.36	0.64	0.61	0.56	0.55	0.54	0.70	0.72	0.62	0.60	0.60	0.73
	75	**0.04**	**0.02**	**0.03**	**0.02**	**0.02**	**0.22**	**0.16**	**0.14**	**0.13**	**0.12**	**0.41**	**0.21**	**0.19**	**0.18**	**0.15**
	125	0.03	0.02	0.03	0.03	0.11	0.20	0.13	0.12	0.10	0.19	0.38	0.19	0.18	0.17	0.23
	175	0.03	0.03	0.03	0.04	0.16	0.19	0.11	0.11	0.11	0.22	0.36	0.18	0.16	0.15	0.25
125	25	0.28	0.29	0.31	0.34	0.65	0.60	0.58	0.57	0.56	0.72	0.71	0.64	0.61	0.61	0.75
	75	0.11	0.11	0.12	0.13	0.36	0.46	0.35	0.32	0.31	0.45	0.60	0.43	0.38	0.37	0.47
	125	**0.03**	**0.02**	**0.02**	**0.02**	**0.02**	**0.19**	**0.11**	**0.08**	**0.07**	**0.07**	**0.36**	**0.17**	**0.13**	**0.12**	**0.09**
	175	0.03	0.02	0.02	0.02	0.05	0.17	0.09	0.08	0.07	0.10	0.33	0.15	0.12	0.11	0.13
175	25	0.27	0.29	0.32	0.33	0.65	0.61	0.57	0.57	0.57	0.69	0.72	0.62	0.63	0.62	0.73
	75	0.12	0.12	0.12	0.15	0.41	0.44	0.39	0.37	0.35	0.48	0.59	0.46	0.42	0.41	0.52
	125	0.07	0.06	0.07	0.08	0.22	0.36	0.27	0.22	0.22	0.28	0.52	0.33	0.27	0.27	0.31
	175	**0.02**	**0.01**	**0.02**	**0.01**	**0.02**	**0.16**	**0.07**	**0.06**	**0.06**	**0.04**	**0.34**	**0.14**	**0.11**	**0.10**	**0.06**

Notes: See notes to Table 1. The test statistic used takes the form of the Wald-statistic described in the fourth bullet of Section 3 (page 8). Consequently, $\chi^2(2)$ critical values are used for inference (as opposed to the standard Normal critical values used in Table 1). The long-run variance used in the calculation of the test statistic was estimated using the Bartlett kernel with the bandwidth set to 1 in the case of $\tau = 1$ and $\lfloor 4((P-\tau)/100)^{2/9} \rfloor + 1$ otherwise. Note that τ observations were lost because the instrument vector in the test statistic includes a τ-lagged loss differential.

Table 6: Conditional: Fixed DGP and Rolling Model, Wald Form

		$\tau = 1$					$\tau = 3$					$\tau = 12$				
				P					P					P		
R	$\tilde{R}$	25	75	125	175	1000	25	75	125	175	1000	25	75	125	175	1000
25	25	0.08	0.05	0.06	0.08	0.43	0.40	0.29	0.27	0.28	0.57	0.57	0.38	0.36	0.36	0.59
	75	0.04	0.05	0.06	0.08	0.30	0.29	0.20	0.17	0.17	0.36	0.46	0.27	0.24	0.23	0.40
	125	0.05	0.05	0.08	0.09	0.33	0.26	0.18	0.18	0.19	0.38	0.41	0.27	0.26	0.23	0.39
	175	0.04	0.06	0.07	0.11	0.35	0.24	0.19	0.19	0.19	0.38	0.42	0.26	0.24	0.23	0.41
75	25	0.34	0.37	0.31	0.30	0.56	0.73	0.76	0.69	0.62	0.70	0.77	0.78	0.69	0.65	0.68
	75	0.03	0.02	0.02	0.02	0.14	0.23	0.14	0.11	0.09	0.19	0.40	0.23	0.16	0.14	0.22
	125	0.03	0.02	0.02	0.02	0.11	0.20	0.10	0.08	0.07	0.15	0.37	0.18	0.13	0.12	0.19
	175	0.02	0.02	0.02	0.03	0.12	0.17	0.10	0.08	0.08	0.17	0.35	0.15	0.13	0.11	0.19
125	25	0.35	0.47	0.50	0.47	0.65	0.74	0.83	0.83	0.80	0.80	0.76	0.83	0.84	0.80	0.79
	75	0.10	0.10	0.08	0.07	0.17	0.47	0.40	0.32	0.25	0.23	0.58	0.44	0.37	0.30	0.26
	125	0.02	0.01	0.02	0.02	0.07	0.18	0.10	0.08	0.06	0.11	0.35	0.15	0.13	0.11	0.14
	175	0.02	0.02	0.02	0.02	0.07	0.16	0.08	0.06	0.05	0.09	0.32	0.14	0.11	0.09	0.12
175	25	0.34	0.47	0.56	0.59	0.71	0.73	0.82	0.87	0.86	0.84	0.78	0.83	0.87	0.88	0.82
	75	0.11	0.11	0.11	0.10	0.16	0.46	0.49	0.44	0.39	0.26	0.59	0.52	0.51	0.46	0.29
	125	0.06	0.04	0.04	0.04	0.09	0.37	0.25	0.20	0.16	0.13	0.53	0.33	0.26	0.22	0.15
	175	0.02	0.02	0.02	0.02	0.05	0.15	0.08	0.06	0.06	0.07	0.33	0.14	0.11	0.09	0.09

Notes: See notes to Table 5. No values are bolded because the null hypothesis does not hold for any entry.

Table 7: Conditional: Rolling DGP and Rolling Model, Scalar Form

		$\tau = 1$					$\tau = 3$					$\tau = 12$				
		P					P					P				
R	$\tilde{R}$	25	75	125	175	1000	25	75	125	175	1000	25	75	125	175	1000
25	**25**	**0.05**	**0.06**	**0.05**	**0.05**	**0.05**	**0.12**	**0.09**	**0.08**	**0.07**	**0.07**	**0.17**	**0.10**	**0.09**	**0.08**	**0.07**
	75	0.11	0.10	0.08	0.08	0.07	0.18	0.14	0.11	0.10	0.07	0.20	0.15	0.12	0.11	0.09
	125	0.13	0.14	0.13	0.12	0.10	0.22	0.20	0.16	0.15	0.10	0.23	0.20	0.17	0.16	0.12
	175	0.13	0.18	0.16	0.15	0.13	0.23	0.23	0.21	0.17	0.12	0.24	0.24	0.20	0.19	0.15
75	25	0.07	0.09	0.10	0.15	0.49	0.12	0.14	0.16	0.20	0.53	0.18	0.15	0.18	0.20	0.50
	75	**0.06**	**0.05**	**0.05**	**0.04**	**0.05**	**0.13**	**0.10**	**0.08**	**0.08**	**0.06**	**0.17**	**0.11**	**0.09**	**0.08**	**0.08**
	125	0.07	0.07	0.06	0.07	0.05	0.15	0.11	0.09	0.10	0.07	0.18	0.14	0.11	0.09	0.08
	175	0.08	0.08	0.09	0.08	0.06	0.17	0.14	0.13	0.12	0.08	0.20	0.16	0.15	0.13	0.08
125	25	0.07	0.10	0.13	0.17	0.68	0.13	0.15	0.20	0.24	0.71	0.19	0.18	0.20	0.23	0.66
	75	0.06	0.06	0.07	0.08	0.14	0.13	0.10	0.10	0.10	0.16	0.17	0.12	0.11	0.12	0.17
	125	**0.06**	**0.05**	**0.05**	**0.05**	**0.05**	**0.14**	**0.10**	**0.09**	**0.08**	**0.07**	**0.19**	**0.11**	**0.09**	**0.10**	**0.08**
	175	0.07	0.07	0.07	0.05	0.05	0.15	0.11	0.10	0.09	0.07	0.19	0.13	0.12	0.11	0.07
175	25	0.07	0.10	0.14	0.20	0.77	0.13	0.15	0.19	0.25	0.80	0.19	0.18	0.21	0.26	0.74
	75	0.06	0.06	0.07	0.08	0.19	0.14	0.10	0.10	0.11	0.23	0.18	0.12	0.12	0.13	0.25
	125	0.07	0.06	0.05	0.06	0.09	0.14	0.10	0.09	0.10	0.11	0.19	0.13	0.11	0.10	0.12
	175	**0.07**	**0.06**	**0.05**	**0.06**	**0.05**	**0.14**	**0.10**	**0.08**	**0.08**	**0.07**	**0.19**	**0.13**	**0.10**	**0.09**	**0.08**

Notes: See notes to Table 1. For $\tau = 1$, the bandwidth of the Bartlett kernel used to estimate the long-run variance was set to 1. Otherwise, the same sample-dependent bandwidth of $\lfloor 4(P/100)^{2/9} \rfloor + 1$ was used.

Table 8: Conditional: Rolling DGP and Fixed Model, Scalar Form

		$\tau = 1$					$\tau = 3$					$\tau = 12$				
		P					P					P				
R	$\tilde{R}$	25	75	125	175	1000	25	75	125	175	1000	25	75	125	175	1000
25	25	0.16	0.29	0.36	0.43	0.67	0.27	0.36	0.41	0.45	0.66	0.26	0.36	0.40	0.44	0.65
	75	0.16	0.30	0.35	0.38	0.57	0.28	0.36	0.37	0.39	0.58	0.26	0.35	0.39	0.43	0.60
	125	0.17	0.28	0.33	0.35	0.55	0.28	0.35	0.36	0.38	0.52	0.28	0.35	0.38	0.40	0.56
	175	0.17	0.28	0.33	0.33	0.51	0.28	0.35	0.37	0.38	0.49	0.26	0.34	0.38	0.39	0.53
75	25	0.10	0.17	0.27	0.33	0.63	0.20	0.25	0.31	0.37	0.62	0.23	0.27	0.33	0.38	0.64
	75	0.08	0.15	0.21	0.27	0.52	0.17	0.21	0.27	0.31	0.54	0.20	0.22	0.27	0.31	0.55
	125	0.09	0.15	0.21	0.26	0.50	0.18	0.22	0.26	0.30	0.50	0.21	0.24	0.29	0.32	0.51
	175	0.10	0.16	0.23	0.27	0.46	0.18	0.22	0.27	0.31	0.48	0.21	0.25	0.29	0.32	0.50
125	25	0.10	0.16	0.21	0.28	0.60	0.20	0.22	0.27	0.33	0.60	0.23	0.25	0.30	0.36	0.62
	75	0.08	0.11	0.15	0.20	0.49	0.17	0.16	0.20	0.24	0.49	0.20	0.19	0.22	0.26	0.51
	125	0.07	0.10	0.14	0.18	0.45	0.16	0.16	0.20	0.23	0.47	0.20	0.20	0.22	0.26	0.49
	175	0.07	0.11	0.15	0.19	0.44	0.17	0.17	0.19	0.25	0.44	0.19	0.19	0.22	0.26	0.47
175	25	0.10	0.16	0.22	0.26	0.60	0.19	0.22	0.27	0.30	0.61	0.24	0.25	0.28	0.32	0.62
	75	0.08	0.10	0.12	0.15	0.46	0.17	0.15	0.16	0.21	0.49	0.21	0.18	0.20	0.23	0.49
	125	0.07	0.07	0.12	0.14	0.41	0.16	0.13	0.16	0.18	0.45	0.21	0.16	0.18	0.20	0.45
	175	0.07	0.08	0.11	0.14	0.40	0.16	0.14	0.15	0.19	0.42	0.20	0.16	0.17	0.21	0.43

Notes: See notes to Table 1. For $\tau = 1$, the bandwidth of the Bartlett kernel used to estimate the long-run variance was set to 1. Otherwise, the same sample-dependent bandwidth of $\lfloor 4(P/100)^{2/9} \rfloor + 1$ was used. No values are bolded because the null hypothesis does not hold for any entry.

Table 9: Conditional: Rolling DGP and Rolling Model, Wald Form

		$\tau = 1$					$\tau = 3$					$\tau = 12$				
				P					P					P		
R	$\tilde{R}$	25	75	125	175	1000	25	75	125	175	1000	25	75	125	175	1000
25	**25**	**0.04**	**0.02**	**0.03**	**0.02**	**0.02**	**0.27**	**0.19**	**0.19**	**0.16**	**0.11**	**0.40**	**0.23**	**0.19**	**0.18**	**0.14**
	75	0.06	0.05	0.03	0.04	0.02	0.25	0.16	0.13	0.12	0.08	0.36	0.19	0.16	0.14	0.08
	125	0.07	0.08	0.06	0.06	0.04	0.24	0.19	0.14	0.12	0.08	0.36	0.20	0.16	0.14	0.11
	175	0.07	0.10	0.08	0.08	0.05	0.23	0.19	0.17	0.15	0.09	0.34	0.21	0.19	0.17	0.11
75	25	0.03	0.03	0.04	0.06	0.33	0.25	0.18	0.16	0.18	0.42	0.39	0.21	0.22	0.22	0.41
	75	**0.03**	**0.02**	**0.02**	**0.02**	**0.02**	**0.16**	**0.08**	**0.06**	**0.05**	**0.03**	**0.32**	**0.11**	**0.09**	**0.07**	**0.04**
	125	0.03	0.02	0.02	0.02	0.02	0.16	0.08	0.06	0.06	0.03	0.31	0.13	0.09	0.07	0.04
	175	0.03	0.03	0.03	0.03	0.02	0.16	0.10	0.08	0.07	0.03	0.30	0.13	0.11	0.08	0.04
125	25	0.02	0.03	0.05	0.06	0.50	0.23	0.17	0.18	0.18	0.59	0.39	0.23	0.23	0.22	0.56
	75	0.02	0.02	0.02	0.02	0.06	0.16	0.08	0.06	0.06	0.09	0.32	0.12	0.09	0.09	0.09
	125	**0.03**	**0.02**	**0.02**	**0.02**	**0.01**	**0.14**	**0.06**	**0.05**	**0.04**	**0.03**	**0.28**	**0.11**	**0.07**	**0.06**	**0.04**
	175	0.03	0.02	0.02	0.02	0.02	0.14	0.06	0.05	0.05	0.02	0.28	0.11	0.08	0.07	0.03
175	25	0.02	0.03	0.06	0.08	0.59	0.22	0.17	0.17	0.18	0.68	0.38	0.22	0.23	0.25	0.64
	75	0.02	0.02	0.02	0.02	0.09	0.15	0.07	0.06	0.06	0.12	0.31	0.12	0.09	0.08	0.14
	125	0.02	0.02	0.02	0.02	0.03	0.14	0.06	0.05	0.05	0.05	0.28	0.10	0.07	0.07	0.05
	175	**0.02**	**0.02**	**0.02**	**0.02**	**0.01**	**0.13**	**0.06**	**0.04**	**0.04**	**0.03**	**0.28**	**0.10**	**0.07**	**0.05**	**0.03**

Notes: See notes to Table 5.

Table 10: Conditional: Rolling DGP and Fixed Model, Wald Form

		$\tau = 1$					$\tau = 3$					$\tau = 12$				
				P					P					P		
R	$\tilde{R}$	25	75	125	175	1000	25	75	125	175	1000	25	75	125	175	1000
25	25	0.10	0.22	0.29	0.34	0.61	0.35	0.39	0.45	0.46	0.64	0.43	0.41	0.43	0.47	0.63
	75	0.10	0.20	0.25	0.27	0.50	0.30	0.34	0.37	0.38	0.57	0.38	0.34	0.36	0.40	0.56
	125	0.10	0.18	0.23	0.26	0.45	0.28	0.31	0.32	0.35	0.50	0.38	0.32	0.35	0.36	0.52
	175	0.09	0.18	0.22	0.24	0.40	0.27	0.30	0.31	0.32	0.44	0.35	0.30	0.32	0.34	0.47
75	25	0.06	0.11	0.17	0.22	0.54	0.27	0.25	0.30	0.34	0.57	0.41	0.29	0.33	0.37	0.58
	75	0.04	0.07	0.12	0.16	0.43	0.19	0.18	0.22	0.25	0.45	0.33	0.21	0.25	0.28	0.48
	125	0.04	0.07	0.13	0.16	0.40	0.18	0.16	0.21	0.23	0.42	0.33	0.19	0.23	0.27	0.44
	175	0.04	0.08	0.13	0.16	0.37	0.17	0.16	0.19	0.23	0.39	0.31	0.18	0.23	0.26	0.42
125	25	0.05	0.09	0.14	0.19	0.52	0.26	0.22	0.25	0.28	0.55	0.40	0.26	0.29	0.33	0.57
	75	0.03	0.04	0.07	0.10	0.38	0.17	0.13	0.14	0.18	0.41	0.32	0.17	0.17	0.21	0.43
	125	0.03	0.05	0.07	0.11	0.34	0.16	0.11	0.13	0.17	0.36	0.31	0.15	0.16	0.19	0.39
	175	0.03	0.05	0.08	0.11	0.33	0.16	0.11	0.12	0.17	0.36	0.30	0.15	0.16	0.19	0.38
175	25	0.06	0.09	0.13	0.18	0.51	0.24	0.21	0.23	0.26	0.53	0.41	0.27	0.27	0.31	0.55
	75	0.04	0.04	0.06	0.08	0.36	0.16	0.12	0.11	0.15	0.40	0.32	0.16	0.14	0.16	0.40
	125	0.03	0.03	0.05	0.07	0.31	0.14	0.09	0.09	0.12	0.35	0.29	0.13	0.12	0.14	0.36
	175	0.03	0.03	0.05	0.07	0.28	0.15	0.08	0.09	0.11	0.32	0.29	0.13	0.13	0.15	0.34

Notes: See notes to Table 5. No values are bolded because the null hypothesis does not hold for any entry.