Backtesting Value–at–Risk using Forecasts for Multiple Horizons, a Comment on the Forecast Rationality Tests of A.J. Patton and A. Timmermann

Lennart F. Hoogerheide*
Francesco Ravazzolo**
Herman K. van Dijk***

* Dept. of Econometrics, Faculty of Economics and Business Administration, VU University Amsterdam, and Tinbergen Institute;
** Norges Bank;
*** Erasmus School of Economics, Erasmus University Rotterdam; Econometric Institute, Tinbergen Institute; and VU University Amsterdam.

Backtesting Value-at-Risk Using Forecasts for Multiple Horizons, a Comment on the Forecast Rationality Tests of A.J. Patton and A. Timmermann[1]

Lennart Hoogerheide[*], Francesco Ravazzolo[†] and Herman K. van Dijk[‡]

Abstract

Patton and Timmermann (2011, "Forecast Rationality Tests Based on Multi-Horizon Bounds", *Journal of Business & Economic Statistics*, forthcoming) propose a set of useful tests for forecast rationality or optimality under squared error loss, including an easily implemented test based on a regression that only involves (long-horizon and short-horizon) forecasts and no observations on the target variable. We propose an extension, a simulation-based procedure that takes into account the presence of errors in parameter estimates. This procedure can also be applied in the field of 'backtesting' models for Value-at-Risk. Applications to simple AR and ARCH time series models show that its power in detecting certain misspecifications is larger than the power of well-known tests for correct Unconditional Coverage and Conditional Coverage.

Keywords: Value-at-Risk, backtest, optimal revision, forecast rationality.

JEL codes: C12, C52, C53, C58, G32.

---

[*]Department of Econometrics & Tinbergen Institute, Vrije Universiteit Amsterdam

[†]Norges Bank

[‡]Econometric Institute & Tinbergen Institute, Erasmus University Rotterdam & Vrije Universiteit Amsterdam

1 Introduction

Forecast rationality under squared error loss implies various bounds on second moments of the forecasts across different horizons. For example, the mean squared forecast error should be non-decreasing in the horizon. Patton and Timmermann (2011) propose rationality tests based on such restrictions, including interesting new tests that can be conducted *without* having data on the target variable; that is, these tests can be performed by checking only the 'internal consistency' of the 'term structure' of forecasts.

One of their novel tests that is easily implemented and that performs well in Monte Carlo simulations (in the sense that the actual size is equal to the nominal size and that the power is high) considers the hypothesis of optimal forecast revision in the context of a linear regression of the most recent forecast on the long-horizon forecast and the sequence of interim forecast revisions. That is, it considers the following regression

$$\hat{Y}_{t|t-1} = \tilde{\alpha} + \tilde{\beta}_H \, \hat{Y}_{t|t-H} + \sum_{j=2}^{H-1} \tilde{\beta}_j \left(\hat{Y}_{t|t-j} - \hat{Y}_{t|t-j-1} \right) + v_t, \qquad (1)$$

where the null hypothesis of 'rationality' or 'optimal revision' corresponds to the hypothesis

$$H_0 : \tilde{\alpha} = 0 \quad \cap \quad \tilde{\beta}_2 = \ldots \tilde{\beta}_H = 1. \qquad (2)$$

Note that the time of the variable to be predicted is 'fixed' at time t, while the regressors are the forecasts for this time t 'running backwards', made at time $t-1$ to $t-H$.

For a simple interpretation of the hypothesis, we rewrite the optimal revision regression (1) as

$$\hat{Y}_{t|t-1} - \hat{Y}_{t|t-2} = \tilde{\alpha} + \tilde{\gamma}_H \, \hat{Y}_{t|t-H} + \sum_{j=2}^{H-1} \tilde{\gamma}_j \left(\hat{Y}_{t|t-j} - \hat{Y}_{t|t-j-1} \right) + v_t, \qquad (3)$$

with $\tilde{\gamma}_h \equiv \tilde{\beta}_h - 1$ $(h = 2, \ldots, H)$. In (3) the null hypothesis of 'rationality' or 'optimal revision' obviously corresponds to the hypothesis

$$H_0 : \tilde{\alpha} = 0 \quad \cap \quad \tilde{\gamma}_2 = \ldots \tilde{\gamma}_H = 0. \tag{4}$$

One of the attractive properties of this test proposed by Patton and Timmermann (2011) is that it has a clear intuitive interpretation: under the null hypothesis of *'no expected forecast correction'* the last update of the forecast, $\hat{Y}_{t|t-1} - \hat{Y}_{t|t-2}$, does not need to *correct* a bias of $\hat{Y}_{t|t-2}$ $(\tilde{\alpha} = 0)$, or the previous updates $\hat{Y}_{t|t-j} - \hat{Y}_{t|t-j-1}$ $(\tilde{\gamma}_j = 0$ for $j = 2, \ldots, H-1)$, or the long-horizon forecast $\hat{Y}_{t|t-H}$ $(\tilde{\gamma}_H = 0)$.

In this paper we address several points. Our main point is to exploit the fact that no actually observed target variable is required and to extend the analysis of Patton and Timmermann to the case of risk measures such as Value-at-Risk and Expected Shortfall for which we never observe the *true* value. The tests can also be used for volatility or variance measures.

The remainder of this paper is organized as follows. In Section 2 we propose an extension of the optimal revision test of Patton and Timmermann (2011), a novel simulation based procedure for testing the validity of a model for forecasting Value-at-Risk. We introduce two versions of the test: (i) for both the validity of the model and the estimated parameters, and (ii) for the validity of the model allowing for estimation errors in the parameters. We show that the test may involve highly non-Gaussian errors, in which case our simulation based testing procedure still performs well. Applications to simple AR and ARCH time series models show that its power in detecting certain misspecifications is larger than the power of well-known tests for correct Unconditional Coverage and Conditional Coverage. In section 3 we present the issues that one encounters when applying the test to an in-sample window of data for which the model is estimated. Remarks about Bayesian testing of inequalities can be found in Section 4. Section 5 concludes.

2   Backtesting Value-at-Risk using forecasts for multiple horizons: a test for optimal revision

Consider the following example in which the target variable evolves according to a stationary AR(2) process

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \varepsilon_t, \qquad \varepsilon_t \sim iid\, N(0, \sigma^2) \tag{5}$$

with $\phi_0 = 0$, $\phi_1 = 0.5$, $\sigma^2 = 1$. For ϕ_2 we consider several values: $\phi_2 = 0.0, 0.1, 0.2, 0.3$. We estimate a simple AR(1) model, (5) with $\phi_2 = 0$. We simulate 1000 data sets of 1500 observations, where the first 1000 in-sample observations are used for (OLS) estimation of the parameters $\theta = (\phi_0, \phi_1, \sigma^2)'$ and the last 500 out-of-sample observations are used for evaluation of Value-at-Risk forecasts. Define $VaR_{t|t-h}^{95\%}$ as the 5% quantile of the predicted distribution of Y_t at time $t - h$ ($h = 1, 2, \ldots$):

$$VaR_{t|t-h}^{95\%} = \hat{Y}_{t|t-h} + \hat{\sigma}\, \Phi^{-1}(0.05) \qquad \text{with} \qquad \hat{Y}_{t|t-h} = \hat{\phi}_0\, \frac{1 - \hat{\phi}_1^h}{1 - \hat{\phi}_1} + \hat{\phi}_1^h\, Y_{t-h}.$$

These $VaR_{t|t-h}^{95\%}$ take the role of $\hat{Y}_{t|t-h}$ in (3), which thus becomes:

$$VaR_{t|t-1}^{95\%} - VaR_{t|t-2}^{95\%} = \tilde{\alpha} + \tilde{\gamma}_H VaR_{t|t-H}^{95\%} + \sum_{j=2}^{H-1} \tilde{\gamma}_j \left(VaR_{t|t-j}^{95\%} - VaR_{t|t-j-1}^{95\%} \right) + v_t. \tag{6}$$

Our null hypothesis is not

$$H_0: \text{forecast rationality or optimality under squared error loss} \tag{7}$$

but

$$H_0: \text{the estimated model for VaR prediction is correct.} \tag{8}$$

That is, we use the test regression (6) without requiring the assumption of squared error loss. The price for this is that, to the best of our knowledge, one generally has to use simulation from the assumed model to generate the distribution of the F-statistic for the null hypothesis in (4). However, for the AR(1) model with i.i.d. $N(0, \sigma^2)$

errors, the errors in (6) are given by $v_t = \phi_1 \varepsilon_{t-1} \sim i.i.d.\ N(0, \phi_1^2 \sigma^2)$, so that under H_0 the F-statistic has its standard F-distribution. Since $\hat{\sigma}\,\Phi^{-1}(0.05)$ is constant, applying the optimal revision regression test to $VaR_{t|t-h}^{95\%}$ amounts to the test for $\hat{Y}_{t|t-h}$.

Results for the test (with $H = 3$) are presented in the first column of Table 1. Even if the AR(1) model is true ($\phi_2 = 0.0$), then the percentage of rejections (at a nominal size of 5%) is 11.6% (with a numerical standard error of 1.0%). The obvious reason is that there are errors in the parameter estimates. The Monte Carlo simulation by Patton and Timmermann (2011) (with nominal size of 10%) does not suffer from this phenomenon, as they assume that the process *and its parameter values* are known to forecasters.

If we want to test for the validity of the model, taking into account the presence of errors in parameter estimates, then we must adapt (i.e. increase) the critical value. We propose the following method:

Procedure for optimal revision testing taking into account errors in parameter estimates:

step 1. Compute parameter estimates $\hat{\theta}$ in model for observed time series y (e.g. AR(1) model with $\theta = (\phi_0, \phi_1, \sigma^2)'$); generate forecasts $(1, 2, \ldots, H$ steps ahead); compute F-statistic $F(y)$ in optimal revision regression;

step 2. Simulate N (e.g. $N = 1000$) data series $y^{(i)}$ $(i = 1, \ldots, N)$ – with same number of observations as observed time series y – from estimated model with parameters $\hat{\theta}$;

step 3. Compute parameter estimates $\hat{\theta}^{(i)}$ for each simulated data series $y^{(i)}$ $(i = 1, \ldots, N)$;

step 4. Generate forecasts $(1, 2, \ldots, H$ steps ahead) for each estimated model with parameters $\hat{\theta}^{(i)}$ and data $y^{(i)}$ $(i = 1, \ldots, N)$;

step 5. Compute F-statistic $F(y^{(i)})$ $(i = 1, \ldots, N)$ in optimal revision regression for each set of forecasts from step 4;

step 6. Compare $F(y)$ with the desired percentile of the sample of *F-statistics under H_0* $F(y^{(i)})$ $(i = 1, \ldots, N)$ from step 5.

Results for this adapted test (with $H = 3$) are in the second column of Table 1. For $\phi_2 = 0$ the percentage of rejections (at a nominal size of 5%) is 4.6% (with a numerical standard error of 0.7%), so that we have no evidence that the size is distorted.

In order to assess the power of the test, we compare the performance to the well-known Unconditional Coverage (UC) and Conditional Coverage (CC) tests for

6

the 95% and 99% Value-at-Risk; see Kupiec (1995) and Christoffersen (1998). In this example, for the optimal revision regression the test results are the same for each $100(1 - \alpha)\%$ Value-at-Risk with $\alpha \in (0,1)$. The percentage of rejections for $\phi_2 = 0.1, 0.2, 0.3$ is clearly larger for the optimal revision regression than for the UC and CC tests. Intuitively, this makes sense, since the optimal revision regression uses a large set of forecasts for multiple horizons, whereas the UC and CC tests are only based on the limited information in the set of 0/1 variables that indicate whether the predicted Value-at-Risk is exceeded by the actual observation. The nominal size for the UC and CC tests is chosen somewhat larger than 5%, as the discrete distributions of the test statistics do not allow for an exact nominal size of 5%. The nominal size is 5.4% and 5.0% (5.3% and 6.4%) in the UC and CC tests for the 95% VaR (99% VaR). The critical values for the UC and CC tests are computed by simulating 100000 series of i.i.d. 0/1 variables under H_0, as the asymptotically valid χ^2 distributions may be rather poor approximations in finite samples, especially for the CC test.

Next, consider the example in which the target variable evolves according to a stationary ARCH(2) process

$$
\begin{aligned}
Y_t &= \varepsilon_t \sqrt{\sigma_t^2} \qquad \varepsilon_t \sim iid\, N(0,1) \\
\sigma_t^2 &= \phi_0 + \phi_1 Y_{t-1}^2 + \phi_2 Y_{t-2}^2
\end{aligned}
\tag{9}
$$

with $\phi_0 = 0.5$ and $\phi_1 = 0.5$. For ϕ_2 we consider several values: $\phi_2 = 0.0, 0.1, 0.2, 0.3$. We estimate a simple ARCH(1) model, (9) with $\phi_2 = 0$. Again, we simulate 1000 data sets of 1500 observations, where the first 1000 in-sample observations are used for estimation of the parameters ϕ_0, ϕ_1 and the last 500 out-of-sample observations

are used for evaluation of Value-at-Risk forecasts

$$VaR_{t|t-h}^{95\%} = \sqrt{\hat{\sigma}_{t|t-h}^2}\, \Phi^{-1}(0.05) \qquad \text{with} \qquad \hat{\sigma}_{t|t-h}^2 = \hat{\phi}_0\, \frac{1 - \hat{\phi}_1^h}{1 - \hat{\phi}_1} + \hat{\phi}_1^h\, Y_{t-h}^2.$$

Applying the optimal revision regression test to $VaR_{t|t-h}^{95\%}$ (or any other $100(1-\alpha)\%$ VaR with $\alpha \in (0,1)$) amounts to the test for the standard deviation $\sqrt{\hat{\sigma}_{t|t-h}^2}$. In this case we cannot even use the critical value from the standard F-distribution for the first, 'strict' optimal revision test (of validity of the model including the parameter values) for two reasons. First, the regressors in (6) may even have small explanatory power for the regressand if the model is correct. For example, in the ARCH(1) model the regressors have no explanatory power for the regressand in test regression (3) for the variance $\hat{\sigma}_{t|t-h}^2$, but since the VaR is proportional to the standard deviation $\sqrt{\hat{\sigma}_{t|t-h}^2}$ this is not necessarily true. Second, the errors v_t in the optimal revision regression (6) can be substantially non-Gaussian, having a negatively skewed and fat-tailed distribution. The histogram in the top panel of Figure 1 shows the negative skewness of the distribution of the errors v_t for one data set simulated from the ARCH(1) model. This skewness is caused by the negative skewness of the distribution of the regressand $(VaR_{t|t-1}^{95\%} - VaR_{t|t-2}^{95\%})$ in (6); the latter is illustrated by the histogram in the middle panel. The bottom panel shows the reason for the negative skewness of $(VaR_{t|t-1}^{95\%} - VaR_{t|t-2}^{95\%})$: $VaR_{t|t-2}^{95\%}$ is more 'moderate' than $VaR_{t|t-1}^{95\%}$, since $VaR_{t|t-2}^{95\%}$ is closer to the unconditional VaR. Therefore $VaR_{t|t-1}^{95\%}$ is sometimes much more negative than $VaR_{t|t-2}^{95\%}$, whereas it is often slightly less negative. The result is a distribution of $(VaR_{t|t-1}^{95\%} - VaR_{t|t-2}^{95\%})$ that has a positive mode and substantially negative skewness. The small differences between the histograms of the errors v_t and the dependent variable $(VaR_{t|t-1}^{95\%} - VaR_{t|t-2}^{95\%})$ reflect that the regressors in (6) have small explanatory power for the regressand, even though the ARCH(1) model is correct. For these reasons, the actual size may be much larger than the nominal size if we would use the critical value from the F-distribution (e.g. an actual size larger than 50% for a nominal size of 5%). Therefore we require simulation for the critical

value in both versions of the optimal revision test. There is also heteroskedasticity for which we use Weighted Least Squares (WLS), assuming $\text{var}(v_t)$ proportional to $\text{var}(y_{t-1})$ (which seems to be a usable approximation). The aim of WLS is to increase the power of the test; the computation of the critical value by simulation already takes care of the size.

In the first test (of validity of the model including the parameter values) we perform the procedure without step 3, using the 'true' parameters $\hat{\theta}$ (instead of $\hat{\theta}^{(i)}$) of our simulated data series in steps 4 and 5. Results are in Table 2. Again, the percentage of rejections of the first optimal revision test is larger than 5% for $\phi_2 = 0$, reflecting the effect of errors in parameter estimates. For the second optimal revision test we do not have evidence that the actual size deviates from 5%. The optimal revision test again has greater power than the UC and CC tests.

In the optimal revision regression test in the AR(1) model a very wrong value of $\hat{\sigma}$ cannot be detected, since the value of $\hat{\sigma}$ does not affect the F-statistic. The UC and CC tests can detect this, which stresses that the optimal revision regression test should preferably be used *in addition to different tests*.

Table 1: Estimated AR(1) model for simulated data from AR(2) model: percentage of rejections (size or power) at 5% nominal size in optimal forecast revision regression test, and tests for unconditional coverage (UC) and conditional coverage (CC) of Value-at-Risk forecasts. Results are computed for 1000 simulated data sets. Numerical standard errors are given between parentheses.

ϕ_2	95% VaR (or 99% VaR) optimal forecast revision ($H = 3$)				95% VaR				99% VaR			
	H_0 : model is correct including estimated parameters		H_0 : model is correct allowing for estimation error in parameters		UC		CC		UC		CC	
	H_0 : model is correct including estimated parameters											
0.0	0.116	(0.010)	0.046	(0.007)	0.080	(0.009)	0.084	(0.009)	0.081	(0.009)	0.088	(0.009)
0.1	0.490	(0.016)	0.315	(0.015)	0.101	(0.010)	0.090	(0.009)	0.077	(0.008)	0.085	(0.009)
0.2	0.979	(0.005)	0.934	(0.008)	0.105	(0.010)	0.113	(0.010)	0.078	(0.008)	0.080	(0.009)
0.3	1.000	(0.000)	0.999	(0.001)	0.148	(0.011)	0.161	(0.012)	0.097	(0.009)	0.098	(0.009)

Table 2: Estimated ARCH(1) model for simulated data from ARCH(2) model: percentage of rejections (size or power) at 5% nominal size in optimal forecast revision regression test, and tests for unconditional coverage (UC) and conditional coverage (CC) of Value-at-Risk forecasts. Results are computed for 1000 simulated data sets. Numerical standard errors are given between parentheses.

| ϕ_2 | 95% VaR (or 99% VaR) optimal forecast revision ($H = 3$) | | | | 95% VaR | | | | 99% VaR | | | |
	H_0 : model is correct including estimated parameters		H_0 : model is correct allowing for estimation error in parameters		UC		CC		UC		CC	
					H_0 : model is correct including estimated parameters							
0.0	0.081	(0.009)	0.049	(0.007)	0.087	(0.009)	0.068	(0.008)	0.074	(0.008)	0.084	(0.009)
0.1	0.175	(0.012)	0.111	(0.010)	0.104	(0.010)	0.084	(0.009)	0.083	(0.009)	0.092	(0.009)
0.2	0.386	(0.015)	0.295	(0.014)	0.145	(0.011)	0.112	(0.010)	0.103	(0.010)	0.111	(0.010)
0.3	0.495	(0.020)	0.438	(0.016)	0.197	(0.013)	0.160	(0.012)	0.127	(0.011)	0.132	(0.011)
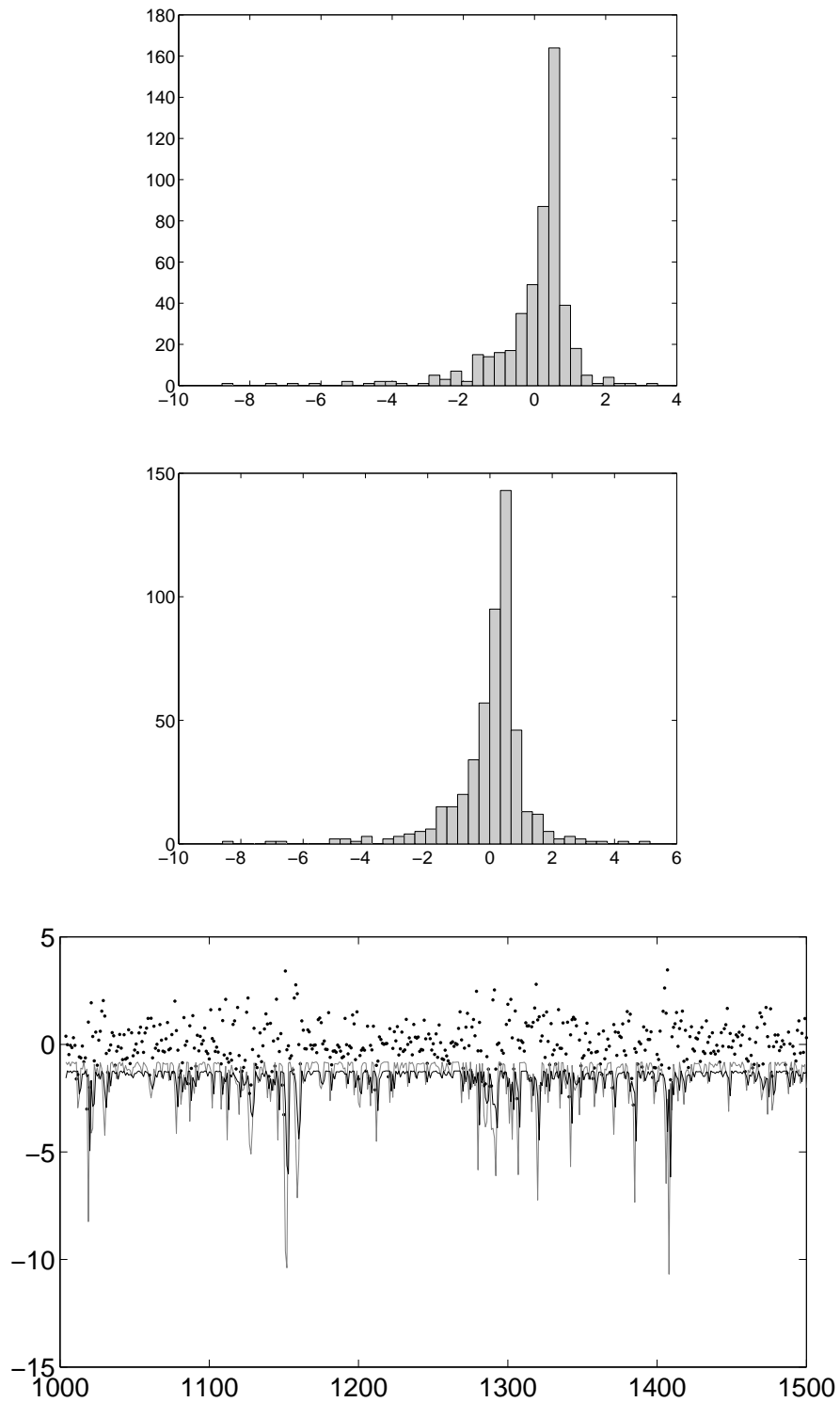
Figure 1: Simulated data set from ARCH(1) model: histograms of error terms v_t [top panel] and regressand $(VaR^{95\%}_{t|t-1} - VaR^{95\%}_{t|t-2})$ [middle panel] in optimal revision test regression (6); graph of simulated data y_t in out-of-sample period (dots), together with $VaR^{95\%}_{t|t-1}$ (grey line) and $VaR^{95\%}_{t|t-2}$ (black line) [bottom panel].

3 The optimal revision test for an in-sample window

If we apply the optimal revision regression test to an *in-sample* window for which the model has been estimated, then a 'generated regressor/regressand problem' implies that the F-statistic does not have the standard F-distribution under H_0, even if the errors v_t are normally distributed. For example, in the AR(1) model we have:

$$\hat{Y}_{t|t-1} = \hat{\phi}_0 + \hat{\phi}_1 Y_{t-1},$$

$$\hat{Y}_{t|t-2} = \hat{\phi}_0 \left(1 + \hat{\phi}_1\right) + \hat{\phi}_1^2 Y_{t-2},$$

$$\hat{Y}_{t|t-1} - \hat{Y}_{t|t-2} = \hat{\phi}_1 \left(Y_{t-1} - \hat{\phi}_0 - \hat{\phi}_1 Y_{t-2}\right).$$

That is, $\hat{Y}_{t|t-1} - \hat{Y}_{t|t-2}$ equals $\hat{\phi}_1$ times the OLS residual, which is obviously perpendicular to the AR(1) model's regressors, the constant term 1, and Y_{t-2}, if we estimate the optimal revision regression (with $H = 2$) for the same window as the parameters ϕ_0, ϕ_1. Then the estimated coefficients ($\hat{\hat{\alpha}}$ and $\hat{\hat{\gamma}}_2$) and F-statistic are exactly equal to 0 for any data series. This reflects that in general the critical values should be smaller if one applies the optimal revision regression test to an in-sample window (or a window that has overlap with an in-sample window).

4   Bayesian testing of inequalities corresponding to forecast rationality

Bayesian inference may be a useful alternative for testing inequalities of (co)variances or coefficients, which is the focus of alternative tests proposed by Patton and Timmermann (2011), especially for small or moderate data samples. Advantages are that no asymptotic approximations need to be used, and that one does not require 'complicated' asymptotic distributions under H_0. A disadvantage is that one needs an explicit model for the distribution, but this may anyway be required for reliable inference in finite samples. We intend to investigate this possibility in further research, simulating from the involved (possibly highly non-elliptical) target distributions by the methods of Hoogerheide, Kaashoek and Van Dijk (2007), Hoogerheide and Van Dijk (2010) and Hoogerheide, Opschoor and Van Dijk (2011).

5   Final remarks

Summarizing, Patton and Timmermann (2011) have proposed a set of interesting and useful tests for forecast rationality or optimality under squared error loss, including an easily implemented test based on a regression that only involves (long-horizon and short-horizon) forecasts and no observations on the target variable. We have discussed an extension, a simulation-based procedure that takes into account the presence of errors in parameter estimates. This procedure can also be applied in the field of 'backtesting' models for Value-at-Risk. Applications to simple AR and ARCH time series models show that its power in detecting certain misspecifications is larger than the power of well-known tests for correct Unconditional Coverage and Conditional Coverage.

References

Christoffersen, P.F. (1998). "Evaluating Interval Forecasts." *International Economic Review* 39(4), 841-862.

Hoogerheide, L.F., Kaashoek, J.F., Van Dijk, H.K. (2007). "On the shape of posterior densities and credible sets in instrumental variable regression models with reduced rank: an application of flexible sampling methods using neural networks." *Journal of Econometrics*, 139(1), 154-180.

Hoogerheide, L.F., Van Dijk, H.K. (2010). "Bayesian Forecasting of Value at Risk and Expected Shortfall using Adaptive Importance Sampling." *International Journal of Forecasting*, 26(2), 231-247.

Hoogerheide, L.F., Opschoor, A., Van Dijk, H.K. (2011). "A Class of Adaptive EM-based Importance Sampling Algorithms for Efficient and Robust Posterior and Predictive Simulation." Tinbergen Institute Discussion Paper TI 2011-004/4.

Kupiec, P.H. (1995). "Techniques for Verifying the Accuracy of Risk Measurement Models." *Journal of Derivatives* 3, 73-84.

Patton, A.,J., Timmermann, A. (2011). "Forecast Rationality Tests Based on Multi-Horizon Bounds." *Journal of Business & Economic Statistics*, forthcoming.