A note about calibration tests for VaR and ES

Luiz Hotta, [1] Carlos Trucíos, [2] Mauricio Zevallos [3]

Resumo

This paper assesses the performance of several calibration tests developed in the recent years to evaluate the forecasting ability of two risk measures commonly used in practice, namely, the value-at-risk (VaR) and expected shortfall (ES). The assessment is made through Monte Carlo experiments on models, with the most popular class of volatility models. The size and power of every test are evaluated under different scenarios considering both normal and Student-t innovation distributions. We also illustrate the application of the calibration tests to two time series of daily financial returns.

Keywords: GARCH; model misspecification; risk measures; expected shortfall

1 Introduction

In recent decades two of the most important risk-management tools to measure capital requirement have been the value-at risk (VaR) and the expected shortfall (ES). These metrics are widely used by financial market investors and regulators. The literature about the statistical properties of these financial risk measures is growing. Accounts of recent developments can be found in Fissler et al. (2016), Fissler and Ziegel (2016), Nolde and Ziegel (2017), Bayer and Dimitriadis (2022) and references therein. A central theme in some of these works is the analysis of calibration tests based on the VaR and/or the ES metrics. For practitioners, this theme is highly relevant because those tests are used for the assessment of risk models. The paper assesses several calibration tests via simulation in a backtesting procedure. The study extends the work of Bayer and Dimitriadis (2022), evaluating the size of various calibration tests and the power of these tests considering the five designs (cases) analyzed in Bayer and Dimitriadis (2022), but in a different aspect. Basically, the difference is the choice of which models are considered as the true and estimated models. The paper is organized as follows. In Section 2, we present a brief description of the calibration tests is presented. The performance of the selected tests is assessed by Monte Carlo simulations discussed in Section 3. Illustrations with real time series returns are presented in Section 4, and concluding remarks are given in Section 5.

2 Calibration tests

Since the VaR and the ES are unobservable, there is no direct way to evaluate their forecasting performance. Nevertheless, calibration tests are a widely used tool to evaluate (indirectly) the forecasting performance of both VaR and ES. See; for instance, Alexander and Dakos (2023) for some recent examples of their use in empirical applications.

In essence, calibration tests are hypothesis tests where roughly the null hypothesis is that the risk measurement procedure is considered *adequate*. Since different authors stress different aspects by which risk measures should be considered *adequate*, there are plenty of calibration tests available in the literature. In this note, we focus on the most popular of them, which are briefly described in Sections 2.1 and 2.2. Hereafter, in all calibration

[1]Departamento de Estatística, Universidade Estadual de Campinas, Campinas, hotta@unicamp.br
[2]Departamento de Estatística, Universidade Estadual de Campinas, Campinas, ctrucios@unicamp.br
[3]Departamento de Estatística, Universidade Estadual de Campinas, Campinas, amadeus@unicamp.br

tests, consider that one-step-ahead VaR and ES are estimated at risk level α at the t-th observation, $t = T+1, ..., T+n_{out}$, with T being the size of the rolling window and n_{out} the size of the out-of-sample period. Denote by $\hat{v}_t$ and $\hat{s}_t$ the estimates of VaR_t and ES_t at a given risk level α, respectively, where, for simplicity, we drop the index α and use the same notation for estimators and estimates.

2.1 VaR calibration tests

Let y_t be the return at time t. Consider the Hit variable , Hit_t, defined by an indicator variable, i.e., $Hit_t = \mathbb{I}(y_t < \hat{v}_t)$ equal to one if $y_t < \hat{v}_t$ and zero otherwise. When $Hit_t = 1$, we say that a *hit* happens at time t, and denote by H the number of times that $y_t < \hat{v}_t$ in the out-of-sample period. Note that when the VaR is correctly estimated, $\{Hit_t\}$ is a sequence of independent random variables with Bernoulli(α) distribution, and then $H = \sum_{t=T+1}^{T+N} Hit_t \sim$ Binomial(N, α). To test the null hypothesis $H_0 : E[H] = n\alpha$, Kupiec (1995) proposed the *Unconditional Coverage* (UC) test. Since the UC test does not account for time dependence of VaR violations, ignoring conditional coverage, Christoffersen (1998) proposed the conditional coverage tesr CC, which also accounts for independence. Another calibration test widely used by practitioners is the dynamic quantile test of Engle and Manganelli (2004). The test is based on the linear regression. Finally, the last test considered is the Gaglianone et al. (2011)´s test based on the α-quantile regression model.

2.2 ES calibration tests

Since the ES is not an elicitable risk measure (Fissler et al., 2016), most calibration tests for ES involve VaR estimates too. One of the most popular calibration tests for the ES is based on the standardized exceedance residuals and was proposed by McNeil and Frey (2000). Let $e_t = (y_t - \hat{s}_t)/\hat{\sigma}_t$, for $t = T+1, ..., T+N$ be the standardized residuals, where $\hat{\sigma}_t$ stands for the estimated volatility at time t and y_t and $\hat{s}_t$ are as previously defined. The exceedance residuals are defined as the set of standardized residuals such that $y_t < \hat{v}_t$.
Under the assumption that the VaR is correctly specified, the mean of the exceedance residuals is zero. Then, the authors propose to test the null hypothesis $H_0 : E(e_t | y_t < \text{VaR}_t) = 0$ against the alternative hypothesis $H_a : E(e_t | y_t < \text{VaR}_t) < 0$. We denote this test as the ER test. We also consider two Wald-type bilateral tests proposed by Nolde and Ziegel (2017) for the null hypothesis $H_0 : E\left(\boldsymbol{W}(\text{VaR}_t, \text{ES}_t, y_t) | \mathcal{F}_{t-1}\right) = \mathbf{0}$. The first test statistic is called conditional calibration test (CoC), and uses no information besides the risk forecast, while the second test is called general conditional calibration test (CoCg). The last four tests are the expected shortfall regression (ESR) tests proposed by Bayer and Dimitriadis (2022). The first two tests, the *Auxiliary ESR Backtest* (aESR) and the *Strict ESR Backtest* (sESR), which are Wald-type test. Finally, the last two test, are a particular case of the *Strict ESR Backtest* and is called the *Intercept ESR Backtest*, one is two-sided and another an one-sided tests; denote them by iESRt and iESRo, respectively.

3 Monte Carlo Simulations

We evaluate the performance of the calibration tests presented in Section 2. The level of the risk measure is chosen to be 2.5%, as suggested by the Basel Accords, and the size of the tests is 5%. We consider the GARCH(1,1) model given by

$$y_t = \sqrt{h_t} z_t \quad \text{and} \quad h_t = \omega + \alpha y_{t-1}^2 + \beta h_{t-1},$$

with $\omega > 0$, $\alpha \geq 0$, $\beta \geq 0$, persistence is $\alpha + \beta < 1$, and z_t is a strict white noise process with zero mean and unit variance. The performance is assessed by backtesting considering a rolling windows scheme of size n_w and out-of-sample evaluation period of

size n_{out}. For each time in the out-of-sample period, one-step ahead forecasts of VaR and ES were calculated. In Section 3.1 we present a simulation study to estimate the true size of the tests under the null hypothesis that the VaR and ES are correct. We also estimate the true critical values in order to compare the size-adjusted power of the backtests; see Bayer and Dimitriadis (2022). The power of the tests is reported in Section 3.2.

3.1 Size of the tests

The true size of the tests described in Section 2 is estimated using simulation. It is similar to the study performed by Bayer and Dimitriadis (2022), but we also consider some calibration tests for the VaR. The data generating process (DGP) is the GARCH model, also studied by Bayer and Dimitriadis (2022). Here, we consider the autoregressive component equal to zero, and change w to have unconditional variance equal to one. We consider innovations with standardized Student-t distribution with 7 d.f. and in model (**??**) we have $\alpha = 0.1$ and $\beta = 0.85$ as in Bayer and Dimitriadis (2022), but $\omega = 0.05$. Table 1 reports the empirical size of the calibration tests for both risk measures at 2.5% risk level and 5% significance level. The results are based on 10,000 replications and out-of-sample sizes, n_{out}, equal to 500, 1000 and 2500 observations, equivalent to two, four and ten years of daily data, respectively. Note that, since the VaR and ES are considered as known, in each replication we only have to simulate the series and use the n_{out} values of returns, VaR and ES. In each replication, we simulate a series considering a burn-in period equal to 500. From Table 1 some features emerge. First, empirical size values decrease and improve as n_{out} increases for all tests except ER and iESRu for both innovations, and for UC and CC for Student-t innovations. Second, in terms of closeness to the nominal size with $n_{out} = 2500$, the best results are obtained using UC and CC tests for normal innovations, and with iESR, eESR, aESR, UC and CC for Student-t innovations. Thus, for Gaussian innovations these results are in line with those of Bayer and Dimitriadis (2022). In summary, since for some of the tests $n_{out} = 1000$ is not enough, we consider $n_{out} = 2500$ a safer value.

3.2 Power of the tests

We evaluate the power of the tests presented in Section 2 by considering the five designs (cases) analyzed in Bayer and Dimitriadis (2022), but in a different aspect. For five design models as the misspecified model, Bayer and Dimitriadis (2022) estimated the power of the tests presented in Section 2.2 when testing: H_0 : The ES is well estimated by one of the misspecified models, versus H_a : The ES is not well estimated, and they estimated the power of the tests when the true DGP is each model given in Section 3.1. We consider a different approach, we test: H_0 : The ES is well estimated by the model given in Section 3.1, versus the same previous alternative hypothesis. We also estimate the power when the true DGP is one of the misspecified models of Bayer and Dimitriadis (2022). In a certain way, this work and that of Bayer and Dimitriadis (2022) are similar and complementary, but our work is more in line with the traditional statistical and econometric approaches, when we consider the power under different “points" of the alternative hypothesis.
Additionally, we also consider the influence of n_{out} on the power of the test, the case of Gaussian distribution for the innovations and the power of the tests for the VaR tests presented in Section 2.1, whereby we have the following hypotheses: H_0 : The VaR is well estimated by the model in Section 3.1, versus H_a : The VaR is not well estimated; and the power is estimated when the true DGP is one of those in Bayer and Dimitriadis (2022)'s misspecified model. The five alternative models cover several issues widely observed in empirical applications: misspecification in the ARCH effect, unconditional variance, per-

sistence, d.f. and risk level. A brief discussion of each design are given below:
Case 1: ARCH effect. We choose $\tilde{\alpha}$ between 0.03 and 0.2, with $\tilde{\beta} = 0.95 - \tilde{\alpha}$, such that the persistence remains the same. Under the null hypothesis $\alpha = 0.1$.
Case 2: Unconditional variance. The unconditional variance is given by $\tilde{\omega}/(1-\tilde{\alpha}-\tilde{\beta})$, and we vary it from 0.01 to 8, without changing the persistence $\tilde{\alpha}+\tilde{\beta}$, i.e., $\tilde{\omega}$ varies from 0.0005 to 0.4 , while $\tilde{\alpha}+\tilde{\beta} = 0.95$. The true unconditional variance is equal to one.
Case 3: Persistence. The persistence of the shocks varies from 0.9 to 0.999, but without changing the unconditional variance. This is done using $\tilde{\alpha} = d\alpha$, $\tilde{\beta} = d\beta$ and $\tilde{\omega}/(1-\tilde{\alpha}-\tilde{\beta}) = \omega/(1-\alpha-\beta)$. The true value is 0.95.
Case 4: d.f. The d.f. of the Student-t innovation varies from 3 to 35 with true value 7.
Case 5: Risk level. In this case the DGP is the one given in Section 3.1 and we estimate the power of the tests when VaR and ES are estimated at risk levels from 0.5% to 5%.
For simplicity we do not present all the results, but since in many tests and cases we need to have $n_{out} = 2500$, Table 2 presents a summary the simulation for this case. We consider that the test "has power" when the power is at least equal to 0.75. The results are for the Gaussian innovation cases, and for the Student-t cases, the power are similar.
The main conclusion is that all VaR tests present high power when the VaR is evaluated at wrong unconditional variance and risk levels, and also with persistence of at least 0.998 for the VQR test and persistence equal to 0.999 for the other tests. Besides that, the VQR test is the only one which presents power when the VaR is evaluated with wrong ARCH parameters, but only when considered equal to 0.02.
For the ES, no tests present power for the ARCH parameter, except the ESR test when $\tilde{\alpha} = 0.02$. For the unconditional variance, CoC, ESR and iESR tests we have large power for any value of the unconditional variance, while ER and C0Cg tests show power for large unconditional variance, and the CoCg test also has power for $\tilde{\sigma}^2 = 0.05$ and, as expected the iESTo test presents power only when the unconditional variance is larger than the DGP unconditional variance. For the persistence, CoC, ESR and iESRo tests present power for large persistence, while ER, CoCg and iESRo tests present no power. For the d.f., except ER, iESRt and iESRo tests, all present power when the VaR and ES are estimated with large d.f. (approximately normal), while the ER and CoC tests show power for d.f. equal to 3. For the risk level error, we have large power for all tests, except the ER test and, as expected, the one-sided iESRo test when it is estimated with lower risk.

4 Illustration

We consider daily returns of Ford Motor Company stock prices and of the EUR/USD exchange rate. For each series, GARCH(1,1) models with standard normal and standardized Student-t innovations are fitted and a rolling window scheme is used to forecasts the one-step-ahead VaR and ES. For the stock and exchange rate return series, the window sizes are 3411 and 2010 days, respectively. We use the last $n_{out} = 1000$ and $n_{out} = 2500$ to evaluate the out-of-sample performance of both risk measures.
Since the 2.5% risk level is of primary importance from a regulatory point of view, we use this risk level, although, obviously other risk levels can also be used. The results are reported in Table 3. For the Ford Stock case, when the Student-t innovation distribution is used, regardless the size of the out-of-sample period, none of the calibration tests reject the null hypothesis at 5% of level. When the normal distribution is used, for both out-of-sample periods the calibration tests for the VaR do not reject the null, but most of the calibration tests for the ES do. When $n_{out} = 2500$ all of the calibration tests for the ES

reject the null and when $n_{out} = 1000$ two out of the three ESR tests do not reject the null. This could be due to the lack of power of those calibration tests, which is in concordance with the simulation results and supports the use of larger out-of-sample periods. For the EUR/USD exchange rate returns, when the Student-t distribution is used, none of the test reject the null hypothesis when $n_{out} = 1000$ but sESR and aESR reject the null when $n_{out} = 2500$. As evidenced in our Monte Carlo experiments, this can be explained by the fact that the power of the test is larger when $n_{out} = 2500$. For normal errors regardless the out-of-sample size, most of the calibration tests of the ES reject the null hypothesis of adequate estimation, but the VaR estimates are not considered inadequate by any test.

5 Concluding Remarks

The size and power of some VaR and ES tests was assessed in backtesting exercises, through simulations with GARCH models as DGP. These models were perturbed in five aspects frequently encountered in empirical applications. The results indicate that to perform backtesting, out-of-sample values of 2500 are safer than 1000. This implies that large sample sizes are needed using daily returns. In terms of performance, the VaR tests only have power when the VAR are evaluated with wrong risk level and unconditional variance, and when the persistence is close to one. In addition, among the ES tests, the best results, across the five perturbations, are obtained for the sESR and aESR tests. They exhibited power for unconditional variance and risk level, and for some values of the rest three perturbations, in situations with very high persistence, large d.f. and small arch effect values. Finally, for the two empirical analyzed series, the use of Student-t innovations is preferred over normal innovations in GARCH models.

References

Alexander, C. and M. Dakos (2023). Assessing the accuracy of exponentially weighted moving average models for value-at-risk and expected shortfall of crypto portfolios. *Quantitative Finance 23*(3), 393–427.

Bayer, S. and T. Dimitriadis (2022). Regression-based expected shortfall backtesting. *Journal of Financial Econometrics 20*(3), 437–471.

Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review 39*(4), 841–862.

Engle, R. F. and S. Manganelli (2004). CAViaR: conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics 22*(4), 367–381.

Fissler, T. and J. F. Ziegel (2016). Higher order elicitability and Osband's principle. *The Annals of Statistics 44*(4), 1680–1707.

Fissler, T., J. F. Ziegel, and T. Gneiting (2016). Expected shortfall is jointly elicitable with value at risk-implications for backtesting. *Risk 5*(1), 8–16.

Gaglianone, W. P., L. R. Lima, O. Linton, and D. R. Smith (2011). Evaluating value-at-risk models via quantile regression. *Journal of Business & Economic Statistics 29*(1), 150–160.

Kupiec, P. (1995). Techniques for verifying the accuracy of risk measurement models. *The Journal of Derivatives 3*(2), 73–84.

McNeil, A. J. and R. Frey (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. *Journal of Empirical Finance 7*(3-4), 271–300.

Nolde, N. and J. F. Ziegel (2017). Elicitability and backtesting: Perspectives for banking regulation. *The Annals of Applied Statistics 11*(4), 1833–1874.

Table 1: Empirical size. GARCH(1,1) models with normal and Student-t with 7 degrees of freedom distributions, and conditional variance $h_t^2 = 0.01 + 0.10r_{t-1}^2 + 0.85h_{t-1}^2$.

	n_{out}	UC	CC	DQ	VQR	ER	CoC	CoCF	sESR	aESR	iESRt	iESRo
Normal	500	0.061	0.037	0.075	0.147	0.030	0.143	0.070	0.041	0.041	0.030	0.005
	1000	0.046	0.044	0.068	0.107	0.029	0.102	0.063	0.036	0.036	0.036	0.010
	2500	0.045	0.046	0.057	0.080	0.033	0.073	0.056	0.032	0.032	0.035	0.018
T_7	500	0.066	0.039	0.080	0.161	0.018	0.172	0.096	0.074	0.074	0.063	0.004
	1000	0.042	0.038	0.070	0.124	0.019	0.123	0.085	0.056	0.056	0.054	0.008
	2500	0.046	0.042	0.057	0.087	0.023	0.083	0.067	0.044	0.045	0.049	0.015

Table 2: Summary of the results of simulation for the Gaussian innovation cases. We consider that the test has power when the power is at least equal to 0.75. *OK* means that the test has power for all alternative values, while *NO* means that does not have power for any alternative value. Otherwise indicate the cases where the test has power. $n_{out} = 2500$.

Tests	Arch effect ($\tilde{\alpha}$)	Unc. Var. ($\tilde{\sigma}^2$)	Persistence ($\tilde{\gamma}$)	D.f. ($\tilde{\nu}$)	Risk level
UC	NO	OK	0.999	NO	OK
CC	NO	OK	0.999	NO	OK
DQ	NO	OK	0.999	NO	OK
VQR	0.02	OK	≥ 0.988	NO	OK
ER	NO	≥ 3	NO	3	NO
CoC	NO	OK	0.999	3 or ≥ 27	OK
CoCg	NO	0.05 or ≥ 3	NO	≥ 21	OK
sESR/aESR	0.02	OK	≥ 0.988	35	OK
iESRt	NO	OK	0.999	NO	OK
iESRo	NO	$>$ true value	NO	NO	0.05

Table 3: Calibration test p-values for the Ford Motor Company and EUR/USD returns. N and T stand for normal and Student-t errors, respectively.

		n_{out}	UC	CC	DQ	VQR	ER	CoC	CoCg	sESR	aESR	iESRt	iESRo
Ford	T	2500	0.487	0.570	0.900	0.687	0.134	0.475	0.060	0.070	0.077	0.349	0.174
	T	1000	0.241	0.317	0.130	0.354	0.386	0.549	0.343	0.110	0.119	0.536	0.268
	N	2500	0.235	0.406	0.860	0.605	0.000	0.000	0.000	0.001	0.001	0.002	0.001
	N	1000	0.056	0.129	0.129	0.798	0.001	0.019	0.009	0.100	0.103	0.048	0.024
EUR/USD	T	2500	0.569	0.134	0.365	0.381	0.359	0.845	0.512	0.041	0.040	0.915	0.457
	T	1000	0.429	0.307	0.822	0.922	0.704	0.620	0.764	0.535	0.555	0.983	0.508
	N	2500	0.848	0.183	0.604	0.608	0.000	0.000	0.000	0.001	0.001	0.003	0.001
	N	1000	0.841	0.489	0.944	0.950	0.001	0.010	0.009	0.066	0.065	0.028	0.014