Statistical properties of duration-based VaR backtesting procedures in finite sample setting

Marta Małecka[1]

Abstract

A dynamic development in the area of VaR estimation and gradual implementation of risk valuation models based on VaR in investment companies stimulate the need for statistical methods of VaR models evaluation. Moreover changes in recommendations of banking supervision institutions, following Basel Accords issued by the Basel Committee, indicate the direction of research, connected with internal VaR models testing, aimed at accurate evaluation of capital requirements. Previous studies have shown that popular VaR independence testing procedure by Christoffersen, Markov tests, exhibits low power, which constitutes a particularly serious problem in case of finite-sample settings. In the paper, as an alternative to the popular Markov test, we presented the overview of the group of duration-based VaR backtesting procedures. We calculated p-values and explored statistical properties of the tests rejecting a non-realistic assumption of infinite sample size. Specifically Monte Carlo test technique was adopted to provide exact tests, in which we replaced asymptotic distributions with simulated finite sample distributions. Monte Carlo study based on GARCH model was also designed to investigate power of the tests. Through the comparative analysis we found that, in the context of power property, the duration-based approach was superior to Markov test.

Keywords*: VaR, VaR backtesting, Markov test, Haas test, TUFF test, Weibull test, gamma test, EACD test*

JEL Classification: C22, C52, D53, G11

AMS Classification: 62M10, 91B84, 62P05

1. Introduction

In the context of business practice, value-at-risk (VaR) measure is by far the most popular approach to risk valuation. Its widening range of applications constantly boosts scientific discussion on various aspects of VaR. There is a parallel discussion in literature on VaR estimation methods and statistical evaluation of VaR models. Commonly used, Markov test by Christoffersen, aimed at evaluating independence in VaR forecasts has been shown to exhibit unsatisfactory power [12]. For practical significance of the independence property, there has been a constant development in statistical testing procedures aimed at detecting serial correlation in VaR violation series. As an alternative to testing the number of exceptions

[1] University of Lodz, Department of Statistical Methods, Rewolucji 1905 r. 41/43, 90-214 Lodz, Poland, marta.malecka@uni.lodz.pl

and working on Markov property assumption, it was proposed to adopt a duration approach, which is based on transformation of the failure process into the duration series.

In 1995 Kupiec [11] presented the concept of the time-until-first-failure test, in which the reverse of no-hit period is treated as an estimate of the success probability in the Bernoulli model. Both this test and its generalization by Haas – the time between failures test of 2001 [9] – are based on the Bernoulli process assumption. Another line of research was based on the assumption of the memory-free exponential distribution, tested against the alternative that involves some wider class of probability distributions [8, 10]. Further approach utilizing exponential distribution properties, was the regression-based exponential autoregressive conditional duration (EACD) test proposed by Engle and Russel in 1998 [6].

The aim of this paper was to provide a revision of independence VaR tests based on durations between VaR exceptions and to present a comparative analysis of their statistical properties. We compared duration-based approach to the broadly used Markov independence test. Statistical properties of all tests were evaluated with the use of Monte Carlo tests technique, which allowed us to obtain the null distribution of tests statistics in finite sample setting. Such technique has an attraction of providing exact tests based on any statistic whose finite sample distribution is intractable but can be simulated [5]. Power properties of the tests were assessed in the simulation study in which GARCH-normal assumption was adopted.

Section 2 of this paper introduces the methodological framework for duration-based testing. Section 3 outlines the Monte Carlo tests procedure, provides details of the simulation study and contains simulation results. The final section summarizes and concludes the article.

2. Duration-based VaR tests

VaR evaluation framework is based on the stochastic process of VaR exceptions:

$$I_{t+1} = \begin{cases} 1, & r_{t+1} < VaR_t(p) \\ 0, & r_{t+1} \geq VaR_t(p) \end{cases}, \qquad (1)$$

where $p-$ given tolerance level, r_t- value of the rate of return at time t, $VaR_t(p)-$ value of the VaR forecast from moment t. In 1998, in order to test for serial correlation, Christoffersen proposed the Markov test, in which he adopted the assumption that the process **Błąd! Nie można odnaleźć źródła odwołania.** forms a part of a Markov chain. The null hypothesis in Markov test, formulated in terms of conditional probabilities of a single-step transition, $H_0 : \pi_{01} = \pi_{11}$, is verified by the statistic

$$LR_{ind} = -2\log\frac{\hat{\pi}_1^{t_1}(1-\hat{\pi}_1)^{t_0}}{\hat{\pi}_{01}^{t_{01}}(1-\hat{\pi}_{01})^{t_{00}}\hat{\pi}_{11}^{t_{11}}(1-\hat{\pi}_{11})^{t_{10}}} \;\square\; {}_{as}\chi^2_{(1)} \tag{2}$$

where $\hat{\pi}_1 = \frac{t_1}{t_0+t_1}$, $t_0 -$ number of non-exceptions, $t_1 -$ number of exceptions, π_{ij} – probability of transition form the state i to the state j, $\hat{\pi}_{01} = \frac{t_{01}}{t_0}$, $\hat{\pi}_{11} = \frac{t_{11}}{t_1}$, $t_{ij} -$ number of transitions form the state i to the state j [3].

By contrast to testing the Markov process assumption, duration-based tests use a transformation of the underlying $\{I_t\}$ process into a duration series $\{V_i\}$ defined as:

$$V_i = t_i - t_{i-1}, \tag{3}$$

where t_i denotes the day of the violation number i. Resting on the assumption that the $\{I_t\}$ series is drawn from the Bernoulli process, Kupiec [11] proposed a *TUFF* test (time until first failure test) that investigates the time of no-hit sequence until the first VaR violation. The reverse of this time constitutes the estimate of the probability of success in the assumed Bernoulli model. The above test was generalized by Hass who, in 2001, introduced a serial correlation of order 1 test, which requires all durations between violations. Haas test statistic is a natural generalization of the *TUFF* test and takes the form:

$$LR_{ind,H} = -2\ln\left[\frac{\alpha(1-\alpha)^{V_1-1}}{p_1(1-p_1)^{V_1-1}}\right] + \sum_{i=2}^{N} -2\ln\left[\frac{\alpha(1-\alpha)^{V_i-1}}{p_i(1-p_i)^{V_i-1}}\right], \tag{4}$$

where $p_i = \frac{1}{V_i}$, V_1 – time until first failure, V_i – time between $(i-1)^{\text{th}}$ and i^{th} violation [7].

An alternative approach to duration testing is to utilize the exponential distribution as the only memory-free random distribution. The null hypothesis of the exponential distribution may be tested against the alternative distribution that allows correlation in the duration series. The tests are based on the LR framework, hence the null model must be nested in the alternative hypothesis. Therefore the alternative family of distributions, in each variant of the test, involves the exponential distribution as a special case.

The alternative distributions that nest the null hypothesis of the exponential distribution, proposed in the literature, involve Weibull and gamma distributions. In case of the Weibull distribution, the pdf takes the form:

$$f_W(V_i, a, b) = a^b b V^{b-1} e^{-(aV)^b} \tag{5}$$

and includes the exponential distribution as a special case for $b=1$. Therefore the null hypothesis takes the form $H_0 : b=1$ and the Weibull test requires fitting the unrestricted Weibull model and its restricted version for $b=1$. Similarly for $b=1$ the exponential distribution is nested in the pdf of gamma distribution:

$$f_\Gamma(V_i, a, b) = \frac{a^b V^{b-1} e^{-aV}}{\Gamma(b)}. \tag{6}$$

As above, in an unrestricted case it is necessary to maximize the gamma loglikelihood function with respect to parameters a and b [4].

The above tests, based on a distribution of durations between VaR violations do not take any account of the ordering of VaR failures, which is considered in the exponential autoregressive conditional duration (EACD) procedure proposed by Engle and Russel [6]. The EACD test verifies the independence of VaR failures utilizing the regression of the durations on their past values:

$$E_{i-1}(V_i) = a + bV_{i-1}. \tag{7}$$

The exponential distribution assumption is also adopted, which gives the conditional pdf function of the duration V_i of the form:

$$f_{EACD}(V_i, a, b) = \frac{1}{a + bV_{i-1}} e^{-\frac{V_i}{a+bV_{i-1}}}, \tag{8}$$

which, for $b=0$, nests the null model with the exponential distribution.

The above tests require computation of the loglikelihood function for the unrestricted and restricted case, which, if we take account of possible presence of censored durations at the beginning and at the end of the series, takes the form:

$$\ln L(V, \Theta) = C_1 \ln S(V_1) + (1 - C_1)\ln f(V_1) + \sum_{i=2}^{N-1} \ln f(V_i) + C_N \ln S(V_N) + (1 - C_N)\ln f(V_N), \tag{9}$$

where C_i takes the value of 1 if the duration V_i is censored and O otherwise, S is the survival function of the variable V_i and N is the number of VaR failures [4].

3. Size and power properties

With regard to practical implementation of the considered tests, which normally involves finite sample setting, we used a Monte Carlo (MC) tests technique. Such technique provides exact tests based on any statistic whose finite sample distribution can be simulated [5]. Following MC tests procedure, we generated $M = 9999$ realizations of the test statistic S_i from the null model and replaced the theoretical distribution of the test statistic F by its sample analogue based on $S_1, ..., S_M$. To generate the $\{I_t\}$ series under the null, we used the Bernoulli distribution with the probability of success p, equal to the assumed level of VaR failure tolerance. Having calculated the survival function:

$$\hat{G}_M(x) = \frac{1}{M}\sum_{i=1}^{M} 1(S_i \geq x) \tag{10}$$

we were able to compute the empirical quantiles of the test statistic distribution. For the test statistic S_0, the corresponding Monte Carlo p-value was obtained according to the formula:

$$\hat{p}_M(S_0) = \frac{M\hat{G}_M(S_0) + 1}{M + 1}. \tag{11}$$

The simulated distributions showed that all tests tend to be oversized in finite samples. The Haas test statistic exhibited the largest discrepancy in the shape of the simulated and theoretical probability density function, which indicates that asymptotic critical values for small samples can be misleading.

For the power comparison we utilized the Monte Carlo simulation technique. The alternative model was obtained by generating return process from the GARCH-normal model and computing VaR estimates from the model with incorrect parameters. The parameter values in the return data generating process: $\omega = 0.000001$, $\alpha = 0.14$, $\beta = 0.85$ were chosen so as to stay in line with real financial process parameter estimates for daily data on stock markets [13]. VaR forecasts were generated from a model with largely increased persistence: $\omega' = 0.000001$, $\alpha' = 0$, $\beta' = 0.99$. Having obtained the VaR violation series and the resulting duration series, we could compute the test statistics and use the Monte Carlo tests technique to evaluate corresponding p-values. Rejection rates under alternative hypothesis were calculated over 10000 Monte Carlo trials [2]. The study was repeated for sample sizes 250, 500,..., 1500.

In the simulation study we rejected cases for which the test was not feasible, this constituted a nontrivial sample selection rule. This was particularly frequent for small samples when no or a very small number of VaR failures occur. Therefore, following raw power estimates, we reported effective power rates, which correspond to multiplying raw power by the sample selection frequency [4].

The comparative analysis results presented in Tables 1 and 2 indicated superiority of all duration-based tests to the Markov test. Finite sample rejection rates showed that for all sample sizes the Haas test exhibited the highest power. In the light of large discrepancy between the simulated and theoretical probability distribution of the test statistic, the Haas test seemed to be very liberal. The simulated distribution is moved to the right off the theoretical shape, hence theoretical quantiles tend to be too small, translating into increased rejection rates. The Haas test application should thus be limited to the analysis carried out with the use of the Monte Carlo test technique, which through simulation exercise, guarantees the exact test size.

	Series length					
	250	**500**	**750**	**1000**	**1250**	**1500**
Markov test	0.106	0.216	0.354	0.436	0.514	0.604
Haas test	0.568	0.818	0.962	0.987	0.997	1.000
Weibull test	0.241	0.582	0.757	0.856	0.919	0.953
Gamma test	0.115	0.455	0.697	0.814	0.893	0.934
EACD test	0.198	0.345	0.448	0.507	0.548	0.593

Table 1 Raw power of the duration-based tests compared to Markov test.

Rejection rates for Weibull and gamma tests were relatively high. In largest sample the power estimates reached levels over 90%. For small sample sizes the Weibull test was slightly superior to the gamma test. In case of smallest examined sample size of 250 observations the rejection rates were very low, under 30%, which is quite striking in the light of business practice and banking supervision requirements, where the series length of 250 observations is typical. Since the presented results showed tests performance in a simulation study, were the difference in parameter values under the null and alternative hypothesis was huge, the application to the real data may yield even poorer power results for small samples.

	Series length					
	250	**500**	**750**	**1000**	**1250**	**1500**
Markov test	0.097	0.213	0.353	0.436	0.514	0.604
Haas test	0.568	0.818	0.962	0.987	0.997	1.000
Weibull test	0.201	0.559	0.752	0.854	0.919	0.953
Gamma test	0.046	0.341	0.644	0.794	0.885	0.931
EACD test	0.131	0.305	0.435	0.502	0.546	0.592

Table 2 Effective power of the duration-based tests compared to Markov test.

The comparison of raw and effective power estimates revealed small sample problem referring to assumptions limiting practical application of the tests. All of them require some number of VaR violations, which in finite sample setting adversely influences the power. Especially for the regression requirements, the EACD test performance was strongly affected in a negative way in smallest samples.

4. Summary and conclusions

This paper explored the family of tests based on durations between subsequent VaR failures and provided insight into statistical properties of duration-based tests in comparison to commonly used Markov test of Christoffersen. Within the duration-based framework we presented the 1995 Kupiec concept of the time-until-first-failure test and its generalization by Haas – the time between failures test of 1998, which are based on the Bernoulli process assumption. Further line of enquiry was to explore procedures based on the assumption of the exponential distribution tested against the alternative involving a wider class of probability distributions. Finally the regression-based approach by Engle and Russel of 1998 was investigated. Statistical properties of all tests were evaluated with the use of Monte Carlo tests technique, which allowed us to obtain the null distribution of tests statistics in finite sample setting. Power properties of the tests were assessed in the simulation study in which GARCH-normal assumption was adopted.

The comparative analysis indicated superiority of all duration-based tests to the Markov test. Finite sample rejection rates were highest for the Haas test. On the other hand the Haas test statistic exhibited the largest discrepancy in the shape of the simulated and theoretical probability density function, which indicated that asymptotic critical values for small samples can be misleading. Rejection rates for Weibull and gamma tests were relatively high, with the Weibull test being slightly superior to the gamma test. In largest sample the power estimates reached levels over 90% by contrast to the case of smallest examined sample size of 250 observations, where the rejection rates were very low, under 30%. The last observation constitutes a powerful argument in a discussion about time series lengths in risk evaluation in business practice and banking supervision requirements.

References

[1] Berkowitz, J., Christoffersen, P., Pelletier, D., 2011. Evaluating Value-at-Risk Models with Desk-Level Data. Management Science 12 (57), 2213-2227.

[2] Białek, J., 2013. Simulation Study of an Original Price Index Formula, Communications in Statistics – Simulation and Computation (in press, DOI: 10.1080/03610918.2012.700367).

[3] Christoffersen, P., 1998. Evaluating Interval Forecasts. International Economic Review 39, 841-862.

[4] Christoffersen, P., Pelletier, D., 2004. Backtesting Value-at-Risk: A Duration-Based Approach. Journal of Financial Econometrics 1 (2), 84-108.

[5] Dufour, J.M., 2006. Monte Carlo Tests with Nuisance Parameters: A General Approach to Finite-Sample Inference and Nonstandard Asymptotics. Journal of Econometrics 133 (2), 443-477.

[6] Engle, R.F., Russel, J.R., 1998. Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data. Econometrica 66 (5), 1127-62.

[7] Fiszeder, P., 2009. Modele klasy GARCH w empirycznych badaniach finansowych. Toruń: Wydawnictwo naukowe uniwersytetu Mikołaja Kopernika.

[8] Gourieroux, C., 2000. Econometrics of Qualitative Dependent Variables, Paul B. Klassen [trans.]. Cambridge: Cambridge University Press.

[9] Haas, M., 2001. New methods in backtesting. Mimeo. Financial Engineering Research Center Caesar, Friedensplatz, Bonn.

[10] Kiefer, N., 1988. Economic Duration Data and Hazard Functions. Journal of Economic Literature 26, 646-679.

[11] Kupiec, P., 1995. Techniques for Verifying the Accuracy of Risk Measurement Models. Journal of Derivatives 2, 174 -184.

[12] Lopez, J., 1999. Methods for Evaluating Value-at-Risk Estimates. FRBSF Economic Review 2, 3-17.

[13] Małecka, M., 2011. Prognozowanie zmienności indeksów giełdowych przy wykorzystaniu modelu klasy *GARCH*. Ekonomista 6, 843-860.

[14] Virdi, N.K., 2011. A Review of Backtesting Methods for evaluating Value-at-Risk. Conference proceedings, Asia-Pacific Business Research Conference “Research for Progress”, Kuala Lumpur, Malaysia, 21-22 February 2011.