STATISTICS IN TRANSITION new series, Autumn 2014
Vol. 15, No. 4, pp. 627–636

DURATION-BASED APPROACH TO VAR INDEPENDENCE BACKTESTING

Marta Małecka[1]

ABSTRACT

Dynamic development in the area of value-at-risk (VaR) estimation and growing implementation of VaR-based risk valuation models in investment companies stimulate the need for statistical methods of VaR models evaluation. Following recent changes in Basel Accords, current UE banking supervisory regulations require internal VaR model backtesting, which provides another strong incentive for research on relevant statistical tests. Previous studies have shown that commonly used VaR independence Markov-chain-based testing procedure exhibits low power, which constitutes a particularly serious problem in the case of finite-sample settings. In the paper, as an alternative to the popular Markov test an overview of the group of duration-based VaR backtesting procedures is presented along with exploration of their statistical properties while rejecting a non-realistic assumption of infinite sample size. The Monte Carlo test technique was adopted to provide exact tests, in which asymptotic distributions were replaced with simulated finite sample distributions. A Monte Carlo study, based on the GARCH model, was designed to investigate the size and the power of the tests. Through the comparative analysis we found that, in the light of observed statistical properties, the duration-based approach was superior to the Markov test.

Key words: 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 market risk valuation. Its increasing 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-chain-based test

[1] University of Lodz, Department of Statistical Methods. E-mail: marta.malecka@uni.lodz.pl.

(Christoffersen, 1998), aimed at evaluating independence in VaR forecasts has been shown to exhibit unsatisfactory power (Lopez, 1999). For practical significance of the independence property, there has been a constant development in statistical testing procedures aimed at detecting various forms of dependence in VaR violation series. As an alternative to testing the number of exceptions 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.

The duration-based approach was primarily motivated by 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 (Kupiec, 1995). Both this test and its generalization in the form of the time-between-failures test (Haas, 2001) were based on the Bernoulli process assumption. Another line of research explored the properties of the memory-free exponential distribution and included the regression-based exponential autoregressive conditional duration test (EACD test, Engle and Russel, 1998). Further approach utilizing the memory-free property was based on testing the assumption of the exponential distribution against the alternative of a wider class of probability distributions (Christoffersen and Pelletier, 2004).

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. Asymptotic probability distributions of the considered tests are known, however when the number of VaR violations is small, which is common in practice, there may be substantial differences between them and their finite sample analogues. Therefore 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 a technique has a great advantage of providing exact tests based on any statistics whose finite sample distribution is intractable but can be simulated (Dufour, 2006). Power properties of the tests were assessed in the simulation study in which GARCH-process assumption was adopted to stay in line with widely recognized facts about financial time series observed in daily intervals.

Section 2 of this paper introduces the methodological framework for duration-based testing. Section 3, dedicated to the simulation study, outlines the Monte Carlo tests procedure, provides details of the Monte Carlo 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 failures:

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

where $p-$ tolerance level, r_t- value of the rate of return at time t, $VaR_t(p)-$ value of the VaR forecast from moment t. Independence tests, based on the VaR failure process, use various forms of the alternative hypothesis. The alternative of the two-state Markov chain was proposed to test for serial correlation (Christoffersen, 1998). The null hypothesis in Christoffersen's Markov test, formulated in terms of conditional probabilities of a single-step transition in the $\{I_t\}$ process, $H_0 : \pi_{01} = \pi_{11}$, is verified by the statistic

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

where $\hat{\pi}_1 = \frac{t_1}{t_0 + t_1}$, t_0- number of non-exceptions, t_1- number of exceptions, $\pi_{ij}-$ probability of transition from 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. State 0 in the above notation is interpreted as non-exception, while 1 represents VaR failure. Under the null, the probability of an exception at time t does not depend on the state of the process at time $t-1$, which means that null hypothesis is equivalent to the iid Bernoulli series.

By contrast to testing the parameter restriction in the assumed Markov chain, 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. The *TUFF* test (time-until-first-failure test), based on the assumption that the $\{I_t\}$ series is drawn from the Bernoulli process, 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 *TUFF* test generalization to the time-between-failures test (Hass, 2001), which requires all durations between violations, examines time-changeability of the Bernoulli process parameter.

The Haas test statistic, being a natural generalization of the *TUFF* statistic, takes the following 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.

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 dependence in the duration series. Similarly to the Markov and Haas test, the proposed exponential distribution tests are based on the LR framework (Domański et al., 2014), where the null model is 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 the case of the Weibull distribution, the pdf takes the form:

$$f_W(v_i) = a^b b v_i^{b-1} e^{-(av_i)^b} \tag{5}$$

and includes the exponential distribution as a special case for $b=1$. Therefore, the null hypothesis takes the form H0: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 gamma distribution, given by the pdf:

$$f_\Gamma(v_i) = \frac{a^b v_i^{b-1} e^{-av_i}}{\Gamma(b)}. \tag{6}$$

As above, in an unrestricted case, it is necessary to maximize the gamma log likelihood function with respect to parameters a and b (Christoffersen and Pelletier, 2004).

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 (Engle and Russel, 1998). 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:

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

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

The above tests require computation of the log likelihood 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 following form:

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

where C_i takes the value of 1 if the duration V_i is censored and 0 otherwise, S is the survival function of the variable V_i and N is the number of VaR failures (Christoffersen and Pelletier, 2004).

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 a technique provides exact tests based on any statistic whose finite sample distribution can be simulated (Dufour, 2006). 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 computed 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 and they do not converge to the nominal test size of 5% with lengthening time series [Tab. 1]. The differences between the empirical and theoretical distribution quantiles were confirmed by the graphical comparison of the simulated and theoretical densities (Fig. 1-5). The Haas and EACD test statistics exhibited the largest discrepancies in the shape of the simulated and theoretical probability density function, which indicated that practical application of these tests shouldn't be based on the asymptotic distributions. The empirical distribution of the Haas test was moved to the right off the theoretical shape, hence theoretical quantiles tended to be too small, translating into increased rejection rates. In the case of the EACD test the empirical distribution lied to the left of the theoretical curve, which gave undersized rejection rates.

Table 1. Empirical size of the duration-based tests compared to Markov test

	Series length					
	250	**500**	**750**	**1000**	**1250**	**1500**
Markov test	0.078	0.084	0.117	0.130	0.125	0.111
Haas test	0.112	0.135	0.183	0.195	0.223	0.199
Weibull test	0.076	0.073	0.084	0.088	0.108	0.114
Gamma test	0.071	0.080	0.090	0.109	0.128	0.157
EACD test	0.008	0.007	0.009	0.011	0.012	0.015

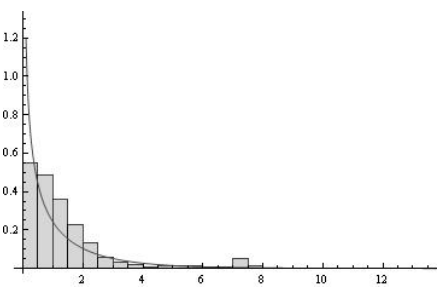

Figure 1. Simulated and theoretical pdf of the Markov test statistics for sample size 250

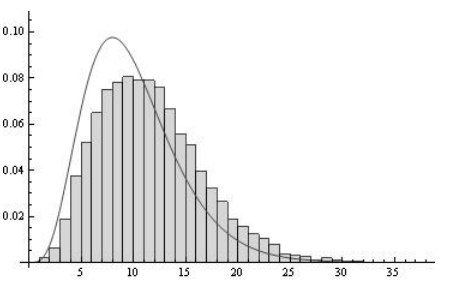

Figure 2. Simulated and theoretical pdf of the Haas test statistics for sample size 250

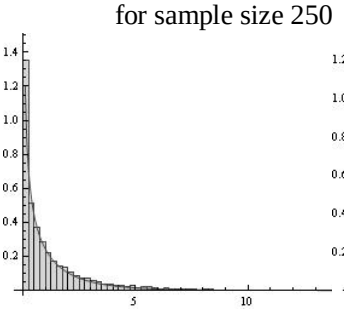

Figure 3. Simulated and theoretical pdf of the Weibull test statistics for sample size 250

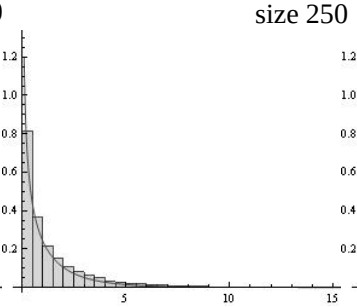

Figure 4. Simulated and theoretical pdf of the gamma test statistics for sample size 250

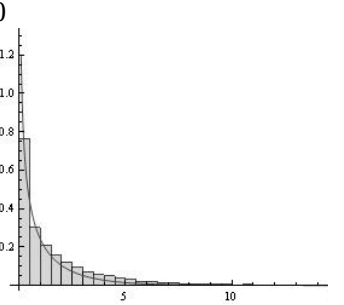

Figure 5. Simulated and theoretical pdf of the EACD test statistics for sample size 250

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 with variance equation of the form $h_t = \omega + \alpha\varepsilon_{t-1}^2 + \beta h_{t-1}$ and computing VaR estimates from the incorrect homoscedastic model. The strength of the correlation in VaR failure series was assessed by the correlation coefficient of the squared returns ρ, whose value was set to 0.1, 0.3 and 0.5 in subsequent variants of the simulation experiment. The parameter values in the return data generating process: $\omega = 0.000001$, $\beta = 0.85$ were chosen so as to stay in line with real financial process parameter estimates for daily data on stock markets (Małecka, 2011). The value of α parameter ensured the required level of ρ (Fiszeder, 2009). VaR forecasts were set to the level of the 0.05 quantile of the unconditional distribution of the return process, which guaranteed the appropriate overall failure rate.

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. The study was repeated for sample sizes 250, 500,..., 1500.

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

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

Test	ρ	Series length					
		250	**500**	**750**	**1000**	**1250**	**1500**
Markov test	0.1	0,082	0,157	0,175	0,250	0,266	0,296
	0.3	0,199	0,423	0,579	0,683	0,749	0,824
	0.5	0,203	0,492	0,611	0,720	0,798	0,857
Haas test	0.1	0,353	0,490	0,622	0,717	0,750	0,843
	0.3	0,390	0,594	0,747	0,834	0,904	0,945
	0.5	0,473	0,662	0,790	0,889	0,924	0,959
Weibull test	0.1	0,064	0,134	0,199	0,318	0,318	0,436
	0.3	0,303	0,679	0,851	0,932	0,968	0,974
	0.5	0,381	0,731	0,878	0,938	0,971	0,985
Gamma test	0.1	0,026	0,051	0,079	0,136	0,144	0,197
	0.3	0,104	0,499	0,745	0,884	0,945	0,967
	0.5	0,120	0,546	0,815	0,915	0,961	0,972
EACD test	0.1	0,135	0,227	0,253	0,279	0,302	0,318
	0.3	0,177	0,358	0,512	0,567	0,621	0,667
	0.5	0,239	0,408	0,503	0,590	0,625	0,676

The comparative analysis of the empirical power (Tab. 2) indicated superiority of the duration-based approach to the Markov test. Finite sample rejection rates showed that for all sample sizes the Haas test exhibited the highest power. It was superior to other tests both for the shortest examined series of 250 observations, which is the minimal series length required for the VaR backtesting by the banking supervision in EU countries, and for the longest series. In all experimental variants, the observed power of the Haas test exceeded 30%. In the case of longest series the empirical power was over 90%. However, in the light of large discrepancy between the empirical and theoretical null distribution, the Haas test application should be limited to the analysis carried out with the use of the Monte Carlo test technique, which guarantees the exact test size.

Comparison of the two procedures based on testing the memory-free property against the Weibull or gamma alternative showed that for small sample sizes the Weibull test outperformed the gamma test. Apart from Haas test, the Weibull approach was another procedure superior to the Markov test. For smallest examined sample size, the observed power of this test was over 30% in two out of three experiment variants. For the largest samples the power estimates reached the levels of over 90%.

From all the considered procedures, including Markov test, the EACD test exhibited the lowest empirical power. This test was also outperformed by all other tests in terms of the test size.

4. Summary and conclusions

The 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 Christoffersen's Markov test of 1998. 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 2001, which are based on the Bernoulli process assumption. Further line of enquiry was the regression-based approach by Engle and Russel of 1998, which utilized the concept of testing the properties of the exponential distribution. Finally we investigated procedures, proposed by Christoffersen and Pelletier in 2004, based on the assumption of the memory-free exponential distribution tested against the alternative involving a wider class of probability distributions. Statistical properties of all tests were evaluated with the use of the 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 based on the GARCH-normal assumption.

The comparative analysis indicated superiority of the duration-based approach to the Markov test. Finite sample rejection rates were the highest for the Haas test. On the other hand, the Haas test statistic exhibited the largest discrepancy in the shape of the empirical and theoretical probability density function, which

indicated that asymptotic critical values for small samples can be misleading. Rejection rates for Weibull tests were higher than for the gamma procedure, also based on checking the memory-free property, and this test was the second duration-based procedure superior to the Markov test. The EACD test was outperformed by all other procedures in terms of both test size and power.

Acknowledgments

The research was supported by the Polish National Science Centre grant DEC-2013/11/N/HS4/03354.

REFERENCES

BERKOWITZ, J., CHRISTOFFERSEN, P., PELLETIER, D., (2011). Evaluating Value-at-Risk Models with Desk-Level Data. Management Science 12(57), 2213–2227.

CHRISTOFFERSEN, P., (1998). Evaluating Interval Forecasts. International Economic Review 39, 841–862.

CHRISTOFFERSEN, P., PELLETIER, D., (2004). Backtesting Value-at-Risk: A Duration-Based Approach. Journal of Financial Econometrics 1(2), 84–108.

DOMANSKI, CZ., PEKASIEWICZ, D., BASZCZYNSKA, A., WITASZCZYK, A., (2014). Testy statystyczne w procesie podejmowania decyzji. Wyd. UŁ, Łódź.

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.

ENGLE, R. F., RUSSEL, J. R., (1998). Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data. Econometrica 66(5), 1127–62.

FISZEDER, P., (2009). Modele klasy GARCH w empirycznych badaniach finansowych. Wydawnictwo naukowe uniwersytetu Mikołaja Kopernika, Toruń.

HAAS, M., (2001). New methods in backtesting. Mimeo. Financial Engineering Research Center Caesar, Friedensplatz, Bonn.

KUPIEC, P., (1995). Techniques for Verifying the Accuracy of Risk Measurement Models. Journal of Derivatives 2, 174–184.

LOPEZ, J., (1999). Methods for Evaluating Value-at-Risk Estimates. FRBSF Economic Review 2, 3–17.

MAŁECKA, M., (2011). Prognozowanie zmienności indeksów giełdowych przy wykorzystaniu modelu klasy GARCH. Ekonomista 6, 843–860.