Parametric hypothesis testing in evaluation of expected shortfall models

Marta Małecka[1]

Abstract

Introduced on the turn of 21st century, the axiomatic risk theory has developed around the notion of a coherent risk measure. In recent literature much attention has been given to the ES (expected shortfall) measure, which fulfils the set of coherency axioms and offers an important extension to the VaR model. As the relevant distribution is unknown, statistical evaluation of the ES model cannot use the natural measure of discrepancy between estimated and empirical ES. Instead hypothesis testing uses the regression approach, the saddlepoint technique or the goodness of fit of the truncated return density. The study presents parametric methods of statistical inference connected with ES measure and, through the simulation study, gives the comparison of the size and power of the considered tests. In order to reflect the stylized facts about real financial processes, simulation experiments are based on the GARCH processes. The power evaluation includes both homoscedastic and heteroskedastic models with incorrect variance parameters.

Keywords*: expected shortfall, test size, test power*

JEL Classification: C22, C52, D53

1. Introduction

The development of the axiomatic risk theory and the inception of the concept of a coherent risk measure, on the turn of the 21st century, gave impetus to introducing new risk models. In particular the models were based on the notion of expected shortfall (ES) whose idea is to inform about the possible loss in case of extreme events. The wide variety of ES-based risk models created the need for relevant testing procedures. In the general case, the distribution of a sample average of extreme observations is unknown, thus classic statistical methods are unfeasible for ES value testing. Since scarcity of observations is inherent to extreme events, the statistical inference cannot be based on the central limit theorem, which requires large sample size.

Since the beginning of the 21st century several approaches have been proposed for ES model backtesting or ES value verification. The censored normal likelihood function was employed in a test comparing empirical and estimated return distribution tail (Berkowitz, 2001). Circumventing the problem of the unknown distribution, the saddlepoint test technique was introduced, which gives approximate p-values through the Taylor expansion of the

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

moment generating function (Wong, 2008). Finally the regression based approach, using the standard Fischer statistic, was proposed (Christoffersen, 2012).

The aim of the paper was to evaluate statistical properties of available parametric ES-testing procedures. Test assessment included their size and power. The analysis of the test properties was preceded by the overview of statistical inference methods proposed in the literature for ES models. We suggested a modification of a Christoffersen's test, aimed at achieving approximate stationarity of the error term in the linear regression in case of stochastic process with a non-constant variance. The statistical properties of the considered tests were evaluated through the Monte Carlo method. In order to reflect the volatility clustering phenomenon, simulation experiments were based on the GARCH processes.

The paper is comprised of five sections. In the second section we introduced the notion of expected shortfall and presented testing procedures dedicated to ES model evaluation or verification of ES value. We also proposed a modification of the regression-based ES test. The sections three and four present and discuss results of the Monte Carlo study of test size and power. The final section summarizes and concludes the paper.

2. Parametric ES tests

The idea behind ES measure is to give information about the possible loss in case of extreme events. Satisfying the postulates of subadditivity, monotonicity, positive homogeneity and translation invariance (Domański, 2011), it matches the axiomatic risk measure definition (Artzner, Delbaen, Eber & Heath, 1999). Let us consider the random variable X defined on (Ω, F, P), such that $E(\max\{0,-X\}) < \infty$. Let $p \in (0,1)$ be a fixed real number. Then expected shortfall of the variable X, at the level of tolerance p, is defined as

$$ES_p(\mathrm{X}) = -\frac{1}{p}\left(E(X\mathbf{1}_{\{X \le q^p(X)\}}) - q^p(X)\left(P(X \le q^p(X)) - p\right)\right), \tag{1}$$

where $q^p(\mathrm{X})$ is the upper p-quantile of X, given by $q^p(X) = \inf\{x \in \mathbf{R} : P(X \le x) > p\}$ (Acerbi, Taasche 2002). In case when X represents a rate of return and is a real random variable, the above definition is equivalent to $ES_p(X) = E(-X \mid X \le q_p(\mathrm{X}))$.

ES tests proposed in recent literature are dedicated to different statistical hypothesis and thus verify risk models through different aspects. The most restrictive procedure, referred to as exception magnitude test, verifies the goodness-of-fit of the return distribution (Berkowitz 2001). Let us consider the random variable R_t, representing the rate of return at time t,

$t = 0,...,\mathrm{T}$. Let F_{R_t} and $\hat{F}_{R_t}$ denote respectively the distribution of R_t and its estimate, $t = 0,...,\mathrm{T}$. In the exception magnitude test the hypothesis of the form $H_0 : \hat{F}_{R_t} = F_{R_t}$ is verified through the first two moments. The test is based on the transformation

$$Z_t = \Phi^{-1}(\mathrm{U}_t) = \Phi^{-1}(\hat{F}_{R_t}(R_t)) \tag{2}$$

where Φ denotes the normal distribution function. Under the null it holds that

$$(Z_t) \overset{i.i.d.}{\sim} N(0,1), \quad t = 1,2,...,T \tag{3}$$

The transformation (2) ensures the normality of random variables, which allows for the use of the wide range of statistical methods based on normality assumption. Statistical verification of the condition (3) may take various forms and rely on moments, serial-correlation or normality testing. In particular, if we consider the linear regression

$$Z_t - \mu = \rho(Z_{t-1} - \mu) + \ň_t, \quad \ň_t \sim N(0, \sigma^2) \tag{4}$$

the hypothesis $H_0 : \mu = 0,\ \sigma = 1,\ \rho = 0$ may be tested through the likelihood ratio, against the alternative $H_0 : \mu \neq 0,\ \sigma \neq 1,\ \rho \neq 0$. The regression (4) may additionally include higher order serial correlation or other exogenous variables.

In order to design the test for ES model, it is proposed to focus exclusively on distribution tails, thus only extreme observations are used to check the condition (3). Let us define the auxiliary variable

$$Z_t^* = \begin{cases} 0, & \text{gdy } Z_t \geq \Phi^{-1}(p) \\ Z_t, & \text{gdy } Z_t < \Phi^{-1}(p) \end{cases} \tag{5}$$

where p denotes the tolerance level.

The censored loglikelihood for parameters μ, σ, ρ and observations $Z_1^*, Z_2^*, ..., Z_T^*$ is given by

$$\begin{aligned} \ln L(\mu, \sigma, \rho, Z_1^*, Z_2^*, ..., Z_T^*) &= \ln \phi\left(\frac{Z_1^* - \mu}{\sigma/(1-\rho)}\right)\mathbf{1}_{(Z_1^* = Z_1)} + \ln\left(1 - \Phi\left(\frac{\Phi^{-1}(p) - \mu}{\sigma/(1-\rho)}\right)\right)\mathbf{1}_{(Z_t^* = 0)} + \\ &+ \sum_{t=2}^{T} \ln\phi\left(\frac{Z_t^* - \mu - \rho Z_{t-1}^*}{\sigma}\right)\mathbf{1}_{(Z_t^* = Z_t)} + \sum_{t=2}^{T} \ln\left(1 - \Phi\left(\frac{\Phi^{-1}(p) - \mu}{\sigma}\right)\right)\mathbf{1}_{(Z_t^* = 0)} \end{aligned} \tag{6}$$

The likelihood ratio test statistic takes the form

$$LR_B^{ES} = -2(\log L(0,1,0, Z_1^*, Z_2^*, ..., Z_T^*) - \log L(\hat{\mu}, \hat{\sigma}, \hat{\rho}, Z_1^*, Z_2^*, ..., Z_T^*) \tag{7}$$

and is asymptotically $\chi^2_{(3)}$ distributed (Berkowitz, 2001).

Another parametric ES testing procedure is based on the saddlepoint technique, which allows for calculating approximate p-values for the sum of random variables (Wong, 2008). It uses Taylor expansion of moment and cumulant generating functions. Originally this method was proposed for the iid normal series, however, through the transformation (2), it is possible to use it for a general class of stochastic processes.

In opposite to the exception magnitude procedure, which checks the fit of the tail distribution, the saddlepoint test is aimed at verifying the hypothesis about the value of ES. Let us assume that the return variable R_t is iid normally distributed, with the density ϕ and the cdf function Φ and let us denote $R = R_t$, $t = 1,...,T$. In such case we can write $ES_p(R_t) = ES_p(R), \quad t = 1,...,T$ and formulate the relevant hypothesis as $H_0 : ES_p(R) = ES_p^{(0)}(R).$ Let us define the random variable X as

$$P(X \le x) = P(R \le x \mid R < q_p) = \frac{P(R \le x)}{P(R < q_p)}, \quad x < q_p, \tag{8}$$

where q_p is the p-quantile of the standard normal distribution. Thus X represents the rate of return on condition that $R < q_p$ and, if p is a chosen tolerance level, then ES can be expressed as $-E(\mathrm{X})$. As the sample average $-\bar{X}$ is a natural estimator of $ES_p(R)$, it is used as a test statistic S in a ES-value test, i.e. $S = -\bar{X}$.

Statistical inference requires the distribution of S or at least the relevant p-value. Using the cumulant generating function $K_X(t)$ of the variable X, the density function of $\bar{X}$ can be written as

$$f_{\bar{X}}(\bar{x}) = \frac{N}{2\Pi}\int_{-\infty}^{\infty} e^{N(K_X(it)-it\bar{x})}dt, \tag{9}$$

Then

$$P(\bar{X} > \bar{x}) = \int_{\bar{x}}^{q_p} f_{\bar{X}}(u)du = \frac{1}{2\Pi i}\int_{s-i\infty}^{s+i\infty} e^{N(K_X(t)-t\bar{x})}t^{-1}dt, \tag{10}$$

where s is the saddlepoint satisfying $K_{X'}(s) = \bar{x}.$ If s satisfies this condition, it is possible to approximate $P(\bar{X} > \bar{x}),$ which serves as the p-value in the ES-value test. Through the Taylor expansion of the moment generating function it can be shown that

$$P(\bar{X} \le \bar{x}) = \begin{cases} \Phi(\xi) - \phi(\xi)\left(\frac{1}{\eta} - \frac{1}{\xi} + \mathbf{O}(N^{-\frac{3}{2}})\right) & \text{dla } \bar{x} < q_p, \\ 1 & \text{dla } \bar{x} \ge q_p, \end{cases} \tag{11}$$

where $\eta = s\sqrt{NK_{X''}(s)}$, $\xi = sgn(s)\sqrt{2N(s\bar{x} - K_X(s))}$, $\bar{x} \neq \mu_x$ (Lugannani & Rice, 1980). In case of normally distributed variable the saddlepoint s can be obtained as a solution to the equation

$$K_{X'}(t) = \frac{M_{X'}(t)}{M_X(t)} = t - e^{\frac{t^2}{2}} \frac{\phi(q_p) - t}{\Phi(q_p) - t} = \bar{x}. \tag{12}$$

The third ES test proposed in the recent literature uses linear regression to verify the potential of additional available random variables to explain ES value. Let us consider the regression

$$-R_{t+1} - ES_p(R_{t+1}) = a + b\mathbf{X}_t + \eta_{t+1}, \text{ where } R_{t+1} < q_p(R_{t+1}), \tag{13}$$

ε_t are iid for $t = 1,...,T$ and $\mathbf{X}_t$ denotes the set of explanatory random variables available at time t. The test hypothesis is formulated as $H_0 : a = 0$, $H_0 : b = 0$ or jointly as $H_0 : a = b = 0$ and can be verified by the standard Fischer statistic, denoted here as F_{Ch} (Christoffersen 2012).

In the general case, the R_t and $ES_p(R_t)$ variables, $t = 1,...,T$, form stochastic processes, whose distributions change over time. In particular, when R_t represents a rate of return from a financial variable, it is characterized by a time-varying variance, which reflects volatility clustering. In case of time-varying variance, the stationarity assumption about the error term ε_t is not satisfied, which may translate into the discrepancy between the theoretical and empirical distribution of the test statistic. Thus statistical inference based on the regression (13) may involve serious type-one error. To reduce the error, taking account of time-variability of R_t variance, we propose the standardization of the dependent variables in the regression (13) with respect the standard deviation. This procedure requires the estimate $\hat{\sigma}_t$. After the standardization, the underlying regression takes the form

$$\frac{-R_{t+1} - ES_p(R_{t+1})}{\hat{\sigma}_t} = a^* + b^*\mathbf{X}_t + \eta^*_{t+1}, \text{ where } R_{t+1} < q_p(R_{t+1}), \tag{14}$$

and η^*_t is the approximately stationary variable. As in case of F_{Ch}, the modified test F^*_{Ch}, based on the regression (14), uses the Fischer statistic.

3. Test size evaluation

The size and power evaluation experiments were designed in a way that they reflected volatility clustering phenomenon, which hinders volatility prediction and is commonly

regarded as a key issue in risk control. Volatility clustering was represented through inclusion of a GARCH process in the data generating algorithm. For the ES test size assessment the R_t values were generated from the *GARCH(1,1)* process:

$$\begin{aligned} R_t &= \sigma_t Z_t, \quad Z_t \sim N(0,1), \\ \sigma_t^2 &= \omega_1 + \alpha_1 R_{t-1}^2 + \beta_1 \sigma_{t-1}^2, \end{aligned} \tag{15}$$

with parameters $\omega_1 = 0,05, \ \alpha = 0,14, \ \beta_1 = 0,85$ [2]. ES values were calculated as expected values of R_t on the interval $(-\infty, q_p(R_t))$, $t = 1,2,...,T$ [3].

In order to obtain U_t values, $t = 1,2,...,T$, in Berkowitz LR_B^{ES} test, we transformed R_t data according to the formula (2), using the true parameters of the data generating process (15). The explanatory variables in the F_{Ch} and F_{Ch}^* tests included five lags of R_t and past five ES predictions. The tests were conducted for the 5% significance level. The rejection frequencies were computed for sample sizes $T = 250,500,750,1000$ over 10000 replications.

Rejection frequencies obtained under the null for LR_B^{ES} and S tests were close to the nominal level of 5% for all series lengths (Tab. 1). The size estimates computed for the regression based test F_{Ch} and the modified version F_{Ch}^* showed that the proposed standardization allowed for significant reduction of the type-one error. The rejection frequencies for the F_{Ch} test more than doubled the nominal level of 5% (Tab. 1). Moreover the results gave no evidence of convergence to the theoretical Fischer distribution. In case of the modified test F_{Ch}^* the rejection frequencies approximately equalled the nominal level and the shape of the statistic distribution for 250 observation got close to the theoretical distribution function (Fig. 1, 2).

Test	**Series length**			
	250	**500**	**750**	**1000**
LR_B^{ES}	0.052	0.053	0.051	0.053
S	0.047	0.054	0.049	0.052
F_{Ch}	0.121	0.203	0.265	0.309
F_{Ch}^*	0.050	0.050	0.049	0.049

Table 1 Size estimates of ES tests.

[2] The parameter values were fixed on the basis of the initial study for six stock market indices (Malecka, 2011).
[3] More about the estimation of ES value can be found in recent literature (Pietrzyk, 2004, Trzpiot, 2010).

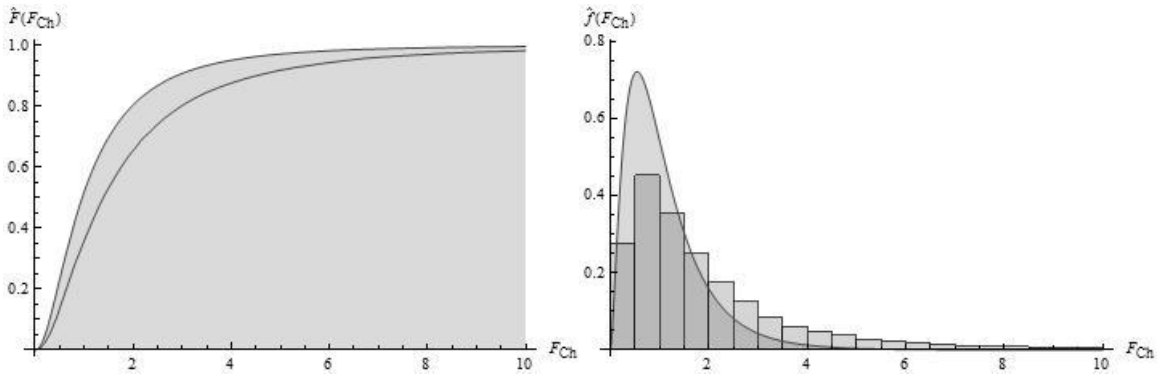

Fig. 1. Empirical and theoretical distribution functions of F_{Ch} test for 250 observations.

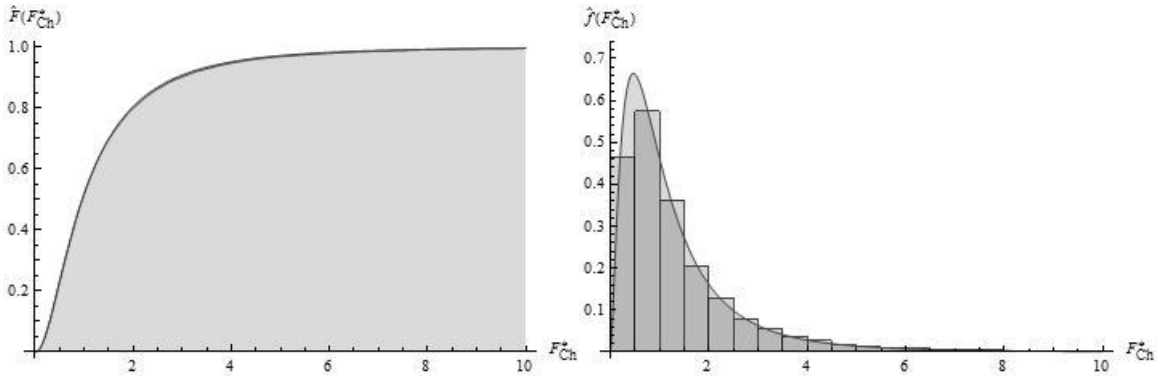

Fig. 2. Empirical and theoretical distribution functions of F^*_{Ch} test for 250 observations.

4. Test power evaluation

For the power comparison, we assumed the data generating process in the form of the *GARCH(1,1)* model given by (15), while ES estimates were obtained from processes with incorrect parameters. We based the power evaluation on two variants of the simulation experiment. In the first stage the ES values were computed from the homoscedastic model with parameters fixed at levels compliant with the unconditional expectation and variance of the data generating process. The resulting failure series was then characterized by the appropriate overall failure rate but the exceedances were serially correlated.

In the second stage the ES values were obtained from the model, which involved time-variability of the return distribution, but with systematically underestimated volatility. We used *GARCH(1,1)* with parameters chosen to obtain the standard deviation on levels of $0.9\sigma_t$, $0.7\sigma_t$ and $0.5\sigma_t$, where σ_t denotes the true parameter value.

Since the testing procedures LR_B^{ES} and F_{Ch}^* presented in the paper are based on asymptotic distributions, the Monte Carlo test technique was employed for the power comparison. Based

on simulated distributions, it provided the empirical quantiles for a given finite sample size and guaranteed the exact test sizes. Hence the power estimates were comparable among different tests. The significance level in the simulation study was set to 5%. The rejection frequencies were computed for sample sizes $T = 250, 500, 750, 1000$ over 10000 replications.

In the first step of the study, the ES values were obtained from the homoscedastic model, incompliant with the time-varying data generating process. The highest rejection frequencies were observed for the LR_B^{ES} test, whose construction is based on the discrepancy between the empirical and theoretical tail of the return distribution. For the series of 250 observations the estimated power of the test nearly reached 70% and it went up to 96% for 1000 observations. The saddlepoint test S rejection frequency for 250 observations was much lower and approached 30%, however it grew to a level of over 60% for the longest sample size.

Test	**Series length**			
	250	**500**	**750**	**1000**
LR_B^{ES}	0.69	0.84	0.93	0.96
S	0.29	0.44	0.56	0.63
F_{Ch}^*	0.08	0.14	0.17	0.21

Table 2 Power estimates of ES tests based on homoscedastic process.

The lowest rejection frequencies were observed for the regression based test F_{Ch}^*. The power estimates for the sample size of 250 observations did not exceed 10%. The results showed a rising tendency with extending the sample, however for the longest series of 1000 observations the estimated power was still approximately as low as 20%.

The second step of the simulation study allowed for a more detailed power comparison, based on experiments where ES estimates were obtained from a model that involved parameter time-variability, however the volatility parameters were undersized. Similarly to the previous experiment, the results showed highest power estimates for the LR_B^{ES} test. The rejection frequencies for this test were over 60% in all experiment variants, for all sample sizes. In case the longest series and standard deviation set to 50% of its true value, the LR_B^{ES} test gave 100% rejection frequency.

The estimated power of the saddlepoint test S, for the series of 250 observations, was between 30% and 90%, depending on the degree of parameter underestimation. There was a

clear growth in the power estimates with lengthening the time series. For 500 observations the rejection frequencies were over 50% and in experiment with standard deviation of 70% of its true value – over 70%. As in case of the LR_B^{ES} test, for the longest series and standard deviation set to 50% of its true value, the S test gave 100% rejection frequency.

Test	σ_t^*	Series length			
		250	500	750	1000
LR_B^{ES}	$0,9\sigma_t$	0.61	0.74	0.82	0.85
	$0,7\sigma_t$	0.62	0.79	0.85	0.90
	$0,5\sigma_t$	0.89	0.97	0.99	1.00
S	$0,9\sigma_t$	0.34	0.51	0.64	0.71
	$0,7\sigma_t$	0.55	0.74	0.85	0.91
	$0,5\sigma_t$	0.87	0.97	0.99	1.00
F_{Ch}^*	$0,9\sigma_t$	0.05	0.05	0.05	0.05
	$0,7\sigma_t$	0.06	0.04	0.05	0.05
	$0,5\sigma_t$	0.06	0.04	0.05	0.05

Table 3 Power estimates of ES tests based on GARCH process with undersized variance.

F_{Ch}^* test rejection frequencies were below 10% for all series lengths. Thus the results, based on the GARCH-type experiments, gave evidence of a very low power of this test against the first order autoregression alternative.

Conclusion

The study presented in the paper was dedicated to evaluation of statistical properties of the parametric ES tests. The results showed therefore that type one errors for the exception magnitude test LR_B^{ES} and saddlepoint test S, assuming series length of at least 250 data, were compliant with the assumed significance level.

For the regression based test F_{Ch} we proposed a modification, which takes account of the time-varying variance of the return distribution. To reduce the type one error, resulting from violation of the stationarity assumption about the error term, we conducted the standardization of the dependent variables. The size estimates computed for the test F_{Ch} and the modified

version F_{Ch}^{*} showed that the proposed standardization allowed for significant reduction of the type-one error.

The power comparison showed that the highest rejection frequencies under the alternatve were observed for the LR_B^{ES} test, whose construction is based on the discrepancy between the empirical and theoretical tail of the return distribution. The saddlepoint test S rejection frequencies were lower, however there was a clear growth in the power estimates with lengthening the time series. The lowest rejection frequencies were observed for the regression based test F_{Ch}^{*}. Despite a rising tendency with increasing the sample, the study gave evidence of a very low power of this test against the first order autoregression alternative, even in case of longest considered time series.

References

Artzner, P., Delbaen, F., Eber, J. M., & Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, *9*(3), 203-228.

Berkowitz, J. (2001). Testing density forecasts with applications to risk management. *Journal of Business and Economic Statistics*, 19, 465-474.

Christoffersen, P. F. (2012). *Elements of Financial Risk Management*. (2nd ed.). Oxford: Elsevier Inc.

Domański, C. (2011). *Nieklasyczne metody oceny efektywności i ryzyka. Otwarte fundusze emerytalne*. Warszawa: Polskie Wydawnictwo Ekonomiczne.

Lugannani, R., & Rice, S. O. (1980). Saddlepoint approximation for the distribution of the sum of independent random variables. *Advanced Applied Probability*, 12, 475-490.

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

Pietrzyk, R. (2004). Szacowanie miary zagrożenia expected shortfall dla wybranych instrumentów polskiego rynku kapitałowego. *Inwestycje finansowe i ubezpieczenia – tendencje światowe a polski rynek*, 1037, 118-127.

Trzpiot, G. (2010). *Wielowymiarowe metody statystyczne w analizie ryzyka inwestycyjnego*. Warszawa: Polskie Wydawnictwo Ekonomiczne.

Wong, W. (2008). Backtesting trading risk of commercial banks using expected shortfall. *Journal of Banking and Finance*, *32*(7), 1404-1415.