Journal of Business Research 89 (2018) 216–222

Contents lists available at ScienceDirect

Journal of Business Research

journal homepage: www.elsevier.com/locate/jbusres

Backtesting an equity risk model under Solvency II☆

Pablo Durán Santomil[a], Luís Otero González[a,*], Onofre Martorell Cunill[b], José M. Merigó Lindahl[c]

[a] Department of Finance and Accounting, Faculty of Economic Science and Business Studies, Universidad de Santiago de Compostela, Avda. Burgo das Nacións s/n, 15704 Santiago de Compostela, Spain

[b] Economía y Empresa, Universitat de les Illes Balears, Carretera de Valldemossa, Km. 7,5, Campus Universitario, EdificioArxiducLluís Salvador, 07071 Palma de Mallorca, Spain

[c] Department of Management Control and Information Systems, University of Chile, Av. Diagonal Paraguay 257, 8330015 Santiago, Chile

ARTICLE INFO

Keywords:
Internal models
Solvency II
Backtesting
Validation
Equity risk
Value-at-risk

ABSTRACT

Backtesting is a technique for validating internal models under Solvency II, which allows for evaluating the discrepancies between the results provided by a model and real observations. This paper aims to establish various backtesting tests and to show their applications to equity risk in Solvency II. Normal and empirical models with a rolling window are used to determine VaR at the 99.5% confidence level over a one-year time horizon. The proposed methodology performs the backtesting of annualized returns arising from the accumulation of daily returns. The results show that even if a model is conservative when tested out of a sample, it may be inadequate when evaluated in a sample, thereby highlighting the problems inherent in the out-of-sample backtesting proposed by the regulator.

1. Introduction

Solvency II is a revision of standards for evaluating the financial situation of European insurers intended to improve risk measurement and control. The main objectives of the new regulation are to protect policyholders and to enhance the stability of the financial systems in the European Union. The dissatisfaction with Solvency I, the influence of Basel II, the Lamfalussy procedure for the elaboration of Community financial regulation, and other international developments[1] have led the European Union to establish a strong need to reform the current solvency system. Like Basel agreements, the reform is based on three pillars that encompass quantitative requirements (Pillar 1), qualitative requirements (Pillar 2) and greater transparency and disclosure (Pillar 3). The starting point of Solvency II is the economic valuation of the whole balance sheet, where all assets and liabilities are valued in accordance with consistent market principles. European companies will have to maintain sufficient financial resources to absorb unexpected losses and to cover the risks inherent in the insurance business. Thus, the regulation requires the maintenance of own funds (i.e., excess of assets over liabilities) and additional own funds[2] classified into three categories (Tier 1, 2, and 3).

Solvency II consists of rules for assessing the financial conditions of insurance companies, with the aim of ensuring that European insurance companies have capital levels adjusted to their assumed real risk. Thus, two capital amounts will be set: (i) economic or risk-based capital (SCR), which is the amount associated with the actual risk taken by the insurer; and (ii) minimum capital (MCR), which is the minimum amount that the insurer must have. The calculation of capital may be carried out through a standard formula or by internal models approved by the regulator. The standard model consists of a set of mathematical formulas that calculate capital requirements, depending on the different risk categories. For the internal models, the introduction of partial models that do not include all risks assumed by an insurer will be permitted. In both cases, the amount obtained correspond to the economic capital that insurance companies should own, measured through the value-at-risk (VaR) with a confidence level of 99.5%.

EIOPA (European Insurance and Occupational Pensions Authority), which from January 1, 2011 has replaced CEIOPS (Committee of

☆ The authors thank professor Txomin Iturralde, UPV university and Luis Ignacio Rodríguez Gil, University of Santiago de Compostela, for their careful reading and suggestions."We gratefully acknowledge financial support provided by the Spanish Ministry of Economy and Competitiveness under the research project with reference ECO2015-71251-R, co-funded by the European Regional Development Fund (ERDF/FEDER) within the period 2014-2020."

* Corresponding author.
E-mail addresses: pablo.duran@usc.es (P. Durán Santomil), luis.otero@usc.es (L. Otero González), onofre.martorell@uib.es (O. Martorell Cunill), jmerigo@fen.uchile.cl (J.M. Merigó Lindahl).

[1] Such developments include the work of the International Association of Insurance Supervisors (IAIS), the International Association of Actuaries (IAA), the International Accounting Standards Board (IASB), and the improvements in the regulation in countries such as Australia, Canada, the USA, Switzerland or the United Kingdom.

[2] Also called supplementary capital. Additional own funds include capital elements that can absorb losses but which are not regarded as having the same strength as own funds.

https://doi.org/10.1016/j.jbusres.2018.01.004
Received 19 June 2017; Received in revised form 2 January 2018; Accepted 3 January 2018
Available online 09 January 2018
0148-2963/

European Insurance Supervisors), is currently responsible for developing the standard model. To perform this task, formulas for calculating the capital for the various risks are proposed, and quantitative impact studies on insurance are performed and used to determine for which average the standard model calibration is adequate. In addition, insurers that wish to use an internal model must submit an application stating that they met certain requirements, among them is the satisfaction of validation rules. It was established that a regular cycle of validation of internal models must be in place, thereby allowing the insurer to prove to the supervisory authorities that the resulting capital requirements are appropriate. Backtesting is a validation technique that allows evaluating the discrepancies between models and actual achievements. Its aim is to improve the quality of an internal model by identifying and analyzing the reasons for the deviations between the actual values and the expected values of a model.

This work aims to establish various backtesting tests and to show their application to equity risk in Solvency II. This work contributes to the existing literature by presenting unpublished evidence on backtesting under the new insurance regulation. The results show that even if a model is conservative when tested out of a sample, it may be inadequate when evaluated in a sample, thereby highlighting the problems inherent in out-of-sample backtesting proposed by the regulator.

Following this introduction, Section 2 reviews the main statistical tests for backtesting. Section 3 presents the data. Section 4 presents the empirical study and the main findings. Finally, Section 5 presents the conclusions as well limitations and suggestions for future research.

2. Backtesting VaR models

Under Solvency II standard model, VaR has been chosen as a widely reported measure in financial markets for providing a simple way to integrate different risks. VaR is the maximum loss that can be expected in a period and the determined confidence interval under normal market conditions. The standard model uses an analytical approach to calculate VaR based on shocks that attempt to measure historic risk. Internal models often determine VaR through simulation methods based on Monte Carlo techniques, which provide greater flexibility by not modelling each risk, thus assuming no particular hypothesis. Formally, VaR is the loss level such that there is a probability p that losses are equal to or greater than Y^*:

$$VaR_p(Y) = Prob(Y \geq Y^*) = p$$

Since the determination of VaR can be performed using alternative models, it is necessary to test the adequacy of the models. This is precisely the objective pursued by backtesting techniques designed to test and validate models are grouped. Backtesting methods aim to analyze the adequacy of internal models based on VaR and mainly consist of statistical tests that attempt to evaluate whether the number of losses exceeding the VaR corresponds to the theoretical percentage.

In implementing a backtesting test, three possible choices that must be resolved: (i) sample versus out of sample, (ii) clean versus dirty, and (iii) dynamic versus static. First, the backtesting can be done in or out of the sample. The so-called *in-sample backtesting* has the problem that it has been fitted to history against past values and therefore may not be expected to be a relatively good adjustment. So and Yu (2006) and Degiannakis and Xekalaki (2007) show that the best models, based on traditional adjustment criteria within the sample, do not lead to the best future estimates. The *out-of-sample backtesting* implies that the model is validated using only observations that occurred after the end of the period of the sample used in the model estimation; thus, a good model within the sample could cause failures when tested out-of-sample. Based on the statements in Article 124 of Directive Solvency II and CEIOPS (2009), backtesting should be carried out outside the sample. Second, *clean backtesting* is a comparative analysis of a portfolio model whose composition remains unchanged. The other possible option is to conduct a *dirty backtesting* on real portfolios of companies that include the results of the operations carried out over time. For insurance companies, the appropriate backtesting is a clean one, because underlying the ideology of Solvency II is stressing the current portfolio within one year using the corresponding worst shocks in 200 scenarios, equivalent to VaR 99.5%. Finally, one can set either a *dynamic* or *static backtesting*. In a dynamic backtesting, the actual values of past returns are used to update the new estimate of VaR. Both methods seem appropriate for the new framework.

Empirical papers that compared different tests in a set of Monte Carlo experiments or using real data are, for example, Campbell (2005), Lehikoinen (2007), Piontek (2009), Virdi (2011), Escanciano and Pei (2012), and Evers and Rohde (2014). Papers on backtesting equity VaR models in Solvency II are very scarce. Otero, Durán, Fernández, and Vivel (2012) find that only Markov Switching Models pass all the tests of four equity indexes at different levels of confidence. Additionally, Moody's Analytics (2013) discusses several aspects of backtesting internal models in the context of 1-year VaR capital models. In addition, Eling and Pankoke (2014) find that the equity capital stress is highly sensitive to the time period considered and to the underlying definition of returns, concluding that applying the equity standard model will lead to systematic deviations from the proposed 99.5% confidence level. Finally, Loois (2015) focuses on the effect of using a rolling horizon to backtest Solvency II VaR models.

2.1. Failures function

The backtesting procedure consists of analyzing the failures induced by the model in relation to the level of optimal failures based on the given level of confidence with which the VaR has been estimated. Therefore, a staple of backtesting is the calculation of the number of times the losses exceed VaR in a given period. In this sense, we build a sequence that takes the value of one (1) if the loss exceeds the VaR and zero (0) otherwise:

$$I_{t+1}(\alpha) = \begin{cases} 1 & if\ x_{t+1} > VaR_t \\ 0 & if\ x_{t+1}.VaR_t \end{cases}$$

where VaR_t is the estimated loss for the moment $t + 1$ using the available information on t; x_{t+1} is the observed loss in $t + 1$, and $I_{t+1}(\alpha)$ is an event indicator of an exception or failure in $t + 1$. The result of applying the function of failures to a given series will be a vector formed by a series of zeros and ones, which indicate whether the losses obtained exceed or fail to exceed the VaR.

In what follows, we discuss some of the main statistical tests for backtesting. The simplest test consists of computing the number of failures and comparing this number with the expected number (i.e., the unconditional coverage test). Other tests take into consideration the hypothesis of independence between failures (i.e., the independence test). There are also joint tests, which attempt to measure the simultaneous fulfilment of two conditions. Multilevel tests simultaneously verify the adequacy of the model for different confidence levels. Finally, other types of tests analyze the relationship between the model estimation results and the actually produced results.

2.2. Unconditional coverage test

A priori, a model will be appropriate when the number of failures match the expected number, which should correspond to a given confidence level. Otherwise, the model does not adequately measure risk. In any case, when conducting a backtesting test for regulatory purposes, it would have less risk to accept models that overestimate the risk than those that underestimate it.

2.2.1. Kupiec's (1995) test or POF

The test proposed by Kupiec (1995) checks whether the number of failures equals α through the Percentage of Fails (POF). When this

failure proportion is different from α, the model either overestimates or underestimates the level of risk. Therefore, the null hypothesis test is:

$$H_0 = \alpha = \hat{\alpha} = \frac{x}{T}$$

where T is the number of observations and x is the number of observed failures. POF statistical test is performed through the ratio between the value of the likelihood function under the null hypothesis and the maximum probability under the alternative hypothesis, which is presented in the denominator:

$$POF = -2\,ln\left(\frac{(1-\alpha)^{T-x}\alpha^{x}}{\left(1-\left(\frac{x}{T}\right)\right)^{T-x}\left(\frac{x}{T}\right)^{x}}\right)$$

Under the null hypothesis POF is X^2distributed with one degree of freedom.

2.2.2. Z-test

For large sample sizes, we can approximate the binomial distribution through a normal distribution. Define a statistical Z as follows:

$$Z = \frac{X-\alpha T}{\sqrt{\alpha(1-\alpha)T}} \approx N(0,1)$$

This test is a variant of the Wald test of the likelihood ratio proposed by Kupiec (1995). An advantage of the Wald test is that it is well defined for the case in which no violation occurs, which is not the case for Kupiec's (1995) test (because the logarithm of zero is undefined).

2.3. Independence test

These tests consider the hypothesis that fails are independent.

2.3.1. Christoffersen's (1998) independence test

The test proposed by Christoffersen (1998) uses a Markovian approach to examine the probability of a failure, which depends on whether another failure has occurred in the previous time period. If the estimated VaR is appropriate, the proportion of failures should be independent of whether a failure has previously occurred. This means that:

$$\frac{n_{00}}{n_{00}+n_{01}} = \frac{n_{10}}{n_{10}+n_{11}}$$

If these proportions differ substantially, the VaR measurement should be questioned. Below we present a Markov's contingency table (Table 1) for the test of independence.

Where n_{00} shows a non-failure at time t and at time $t-1$, n_{10} shows a non-failure at t but a failure at $t-1$ as well as a failure at t and a non-failure at $t-1$; and, finally, n_{11} shows a failure at t that follows another failure at $t-1$. We will denote the total of failures by n_1, and the total non-failures by n_0,. Once we formed a table with the crosstabs, we can calculate the statistical measure as follows:

$$POF_{ind} = -2\,ln\left(\frac{(1-\Pi)^{n_{00}+n_{10}}\Pi^{n_{01}+n_{11}}}{(1-\Pi_0)^{n_{00}}\Pi_0^{n_{01}}(1-\Pi_1)^{n_{10}}\Pi_1^{n_{11}}}\right)$$

where

$$\Pi_0 = \frac{n_{01}}{n_{00}+n_{01}},\ \Pi_1 = \frac{n_{11}}{n_{10}+n_{11}},\ \Pi = \frac{n_{01}+n_{11}}{n_{00}+n_{01}+n_{10}+n_{11}} = \frac{n_1}{n_0+n_1}$$

Under the null hypothesis whereby the model is correct, the statistical POF is X^2 distributed with one degree of freedom.

2.3.2. Berkowitz, Christoffersen, and Pelletier (BCP) (2011) test

The test proposed by Berkowitz, Christoffersen, and Pelletier (2011) focuses on the failed process without average or centered on the α probability, calculated as $I_t - \alpha$, which is a martingale difference in the absence of autocorrelation. Thus, for any variable Z_t, the following is the set of information of an economic agent at time t:

$$E\left[(I_{t+1}-\alpha)\bigotimes Z_t\right] = 0$$

If Z_t is specified as the latest failure without average media, i.e., $(I_t - \alpha)$, we obtain:

$$E\left[(I_{t+1}-\alpha)(I_t-\alpha)\right] = 0$$

Under the null hypothesis, the sequence of failures does not have a first-order autocorrelation. This property is analyzed by Christoffersen's (1998) test. More generally, however, it provides that $Z_t = I_{t-k}$ for any delay $k > 0$:

$$E\left[(I_{t+1}-p)(I_{t-k}-p)\right] = 0$$

Thus, one can check if any of the autocorrelation is significantly different from zero. Under this test, the null hypothesis is that all the autocorrelations are zero against the alternative hypothesis that there exists an autocorrelation that is different from zero:

$$H_0 = \gamma_k = 0, f\ or\ k > 0$$

$$H_1 = \gamma_k \neq 0, \quad \text{for some } k$$

To show this, one can use the Ljung-Box test, which is a joint test that examines whether the first m-autocorrelations are zero. The statistic is defined as $Q = T(T+2)\sum_{(k=1)} m\,\frac{(\gamma_k^2)}{T-k}$,where γ_k is the k-th autocorrelation and T is the number of observations. Q is asymptotically X^2 distributed with degrees of freedom equal to the number of autocorrelations, i.e., m.

2.3.3. CAViaR test

In addition to the past values, one should consider whether the failures can be predicted by including other data such as recent performances, the VaR levels of past periods, and failures related to other confidence levels. Under the null hypothesis, that is equivalent to:

$$E\left[(I_{t+1}-\alpha)g(I_t, I_{t-1},..., z_t, z_{t-1},...)\right] = 0$$

for a function $g(\cdot)$ of the past values of failures and any other variable z_t. For example, following the methodology of Engle and Manganelli (2004), the following regression of order n can be considered:

$$I_t = \alpha + \sum_{k=1}^{n}\beta_1 I_{t-k} + \sum_{k=1}^{n}\beta_2 g(I_t, I_{t-1},..., z_t, z_{t-1},...) + \varepsilon_t$$

For example, if we set $g(I_t, I_{t-1},..., z_t, z_{t-1},...) = (VaR_{t+1} - \overline{VaR})$, one can test whether the β coefficients and the value of α are statistically significant using the typical parameter test of significance. The model that we will analyze later is given by the following equation:

$$(I_{t+1}-\alpha) = \beta_0 + \beta_1(I_t-\alpha) + \beta_2(VaR_{t+1} - \overline{VaR}) + \varepsilon_t$$

These models are referred to in the literature as the CAViaR (Conditional Autoregressive Value at Risk) test. One can test hypothesis that $\beta_0 = 0$ for the coverage conditional, $\beta_1 = \beta_2 = 0$, for

Table 1
Markov's contingency table.

	$I_{t-1} = 0$	$I_{t-1} = 1$	
$I_t = 0$	n_{00}	n_{10}	$n_{00} + n_{10}$
$I_t = 1$	n_{01}	n_{11}	$n_{01} + n_{11}$
	$n_{00} + n_{01}$	$n_{10} + n_{11}$	N

unconditional coverage or the joint test $\beta_0 = \beta_1 = \beta_2 = 0$.

2.4. Joint tests

An accurate measure of VaR should have properties of unconditional coverage and independence. In this sense, the tests that examine both properties at the same time identify measures of VaR that are deficient for failing either of the two properties. The Markov test, proposed by Christoffersen (1998), can be extended to a joint test of both properties. Thus, if the VaR measure also includes ownership of unconditional coverage, then these ratios should also match the total failure rate (α):

$$\frac{n_{00}}{n_{00}+n_{01}} = \frac{n_{10}}{n_{10}+n_{11}} = \frac{n_{00}+n_{10}}{n_{00}+n_{01}+n_{10}+n_{11}} = \alpha$$

Therefore, the whole Markov test measures whether there is any difference in the likelihood of a failure, conditioned on whether or not there has been a previous failure and determines simultaneously whether each of these proportions is significantly different from α. While these tests may seem more appropriate, since both evaluate both properties simultaneously, they have the least capacity to detect measurements of VaR that only one of the two properties do not fulfil (Campbell, 2005). The most commonly used test is obtained by adding Christoffersen's (1998) statistical test of independence to the statistic of Kupiec's (1995) test, which is X^2 distributed with two degrees of freedom.

2.5. Tests based on VaR multiple levels

The above tests only analyze the adequacy of VaR for a given level of confidence. However, an accurate measure of VaR should be valid for any level of confidence. That is, if the calculation of VaR is adequate, 99% VaR should be exceeded in 1% of the cases, 95% VaR in 5%, and so on. In addition, failures that occur within a given level should be independent of those introduced under other levels of trust.

2.5.1. Pearson's Q test

An option proposed by Campbell (2005), called the Pearson test—applicable to multiple levels VaR—is to analyze the behavior of risk measures for a predetermined range of values of α. This test is constructed as follows:

1. We define the unit interval into several subintervals, for example: (0, 0.01), (0.01, 0.05), (0.05, 0.10), and the remainder (0.10, 1).
2. We count exceptions that occur in each interval so that an exception in the first interval exceeds 99% VaR.
3. Finally, the Q test that compares the above frequencies with actual theoretical frequencies is performed:

$$Q = \sum_{i=1}^{k} \frac{(N_{(l_i,u_i)} - N(u_i - l_i))^2}{N(u_i - l_i)}$$

where $N_{(l_i,u_i)}$ is the number of exceptions in a given interval, N is the number of temporary periods used to build the test, and u_i and l_i are the upper and lower values of each interval, respectively. If the model is appropriate, the test is approximately X^2 distributed with one degree of freedom.

2.6. Correlation test

Another type of test is to analyze the correlation between the VaR estimates and the magnitude of the returns. The traditional backtesting tests that evaluate the performance of any risk measure focus on analyzing the level of the coverage provided, but do not address the efficiency of these measures. An appropriate measure of risk has to be not only conservative enough, i.e., provide adequate coverage, but should also be closely related to the exposure of the portfolio. A measure of risk that is conservative but inefficient tends to overestimate risk in periods of low market volatility. The simplest test is to assess the relationship between the VaR estimates and the value of the returns. In this sense, it would be advisable for large VaR figures to be accompanied by large negative returns, whereas small VaR calculations must be associated with small, negative, or positive returns.

Given that the assumption on the normality of returns is not met in many cases, it is useful to consider a test that requires no assumption about the distribution of the two series. Thus, the correlation coefficient tests of Spearman's Rho or Kendall's Tau are used. Let R_t be the returns series and VaR_t the VaR series, and let the coefficients of the rank correlation be the correlation coefficient ordinate variables. The Kendall's Tau statistical analyzes the agreements and disagreements between pairs of data points. If the p-value of any of these tests is lower than its level of acceptance (normally 5%), the null hypothesis that the variables are statistically independent can be rejected and the alternative hypothesis that they are related can be accepted.

3. Data

In the backtesting of internal models, reached VaR levels should be compared with returns obtained in the market. In mathematical terms, the VaR with a level of confidence α is defined as a quantile of the distribution of profit and loss. Assuming that asset returns are normally distributed, the VaR is calculated as follows:

$$VaR_\alpha = \mu + q(1-\alpha)\sigma$$

where q $(1 - \alpha)$ reflects the quartile of the standard normal distribution with the selected level of confidence, the standard deviation (σ), and the average return (μ). In the case of using a time series model for the conditional mean or the conditional volatility of returns, the formula for VaR, assuming that the residuals are normally distributed, will be:

$$VaR_\alpha = \mu_t + q(1-\alpha)\sigma_t$$

The VaR methodology was developed by JP Morgan in the 90s with a focus on the banking industry. However, extrapolation of this methodology to the insurance industry, which is used to medium and long-term time horizons, is questionable. Panning (1999), Dowd, Blake, and Cairns (2004), and Fedor (2007) discuss various problems of VaR when applied to the insurance sector. These studies show that the estimated long-term VaR is more difficult to obtain than the estimated short-term VaR. This is due to the difficulty of predicting long-term volatility; thus, as Christoffersen, Diebold, and Schuermann (1998) and Christoffersen and Diebold (2000) state, the ability to forecast volatility appears to be rapidly reduced as the horizon grows, and seems to largely disappear beyond ten or fifteen day time horizons. Dowd et al. (2004) proposed a simple technique of applying long-term average values of the parameters, whereas Fedor (2007) compares the different VaR time horizons arising from the extrapolation of a one-day VaR by using the rule of square root with several models (i.e., historic, covariance, and Monte Carlo based on the normal distribution).

In the financial sector, which uses a short-term temporary horizon, it is usually assumed that the mean (μ_t) is zero. This option, when applied to long-term periods, would underestimate VaR if the mean is negative (i.e., an economic decay period) and, conversely, overestimate VaR in the event that the average would be possible. Possible options that may address the problems that arise in long-term horizon models are: (i) do not use average term, (ii) use the historical term, or (iii) determine the average based on the expected returns, which require the so-called expert judgment.

Insurance companies normally use monthly or quarterly data, since in many cases the time horizon of their liabilities is decades. In this context, if a portfolio has a monthly VaR of one monetary unit (m.u.) with a confidence level of 99.5%, it means that the portfolio has a 0.5% chance of suffering a loss of at least one m.u. at the end of the month

under normal market conditions. If the portfolio experienced a greater loss of one m.u. at the end of the month, it seems a priori that the model does not properly consider VaR. To achieve formal results, we should perform a monthly backtesting. With a confidence level of 99.5%, we could expect to have a failure over the next 16.67 years. This does not seem feasible under Solvency II, as backtesting should be performed at least annually. The option of using daily data with a temporary horizon of daily backtesting, as used in the banking sector, does not seem appropriate either; this is because, although a model that produces so many daily failures would lead to the rejection of the model, the model test may not be excessive in annual terms, which is how capital charges should be calculated under Solvency II. In addition, the option of calculating daily VaR and using the so-called square root rule for calculating VaR within a year would produce unrealistic results (see Dowd et al., 2004).

The above discussion leads us to apply another method of selecting the daily logarithmic returns and proceed to accumulate them based on 252 daily observations to obtain annual returns (i.e., the rolling window method). To do this, we have selected the IBEX35 index, which accumulated annual returns from January 1996 to the end of October 2013. The objective of this work is to show the application of the above analyzed tests to backtesting of possible models. The analyzed models will be normal model, in which the parameters are re-estimated with a rolling window based on the last 252 observations, or an empirical model or historical model, which calculates VaR as historical loss percentile of the past 252 observations. The choice of these models is justified by the fact that the normal model was the basis of the models used by the CEIOPS up to QIS4, whereas the empirical model underlies the QIS5 studies. EIOPA has used in recent studies the rolling window method for annual returns obtained from daily ones. However, in contrast to EIOPA, the dampener effect for equities in the internal model has not been taken into account, which means that if the stock market is falling, insurance companies are required to obtain lower capital charges. The logic of the dampener effect lies in the attempt to assure that insurance companies do not undo their positions in the equity markets during financial crises, as this would create a pro-cyclical effect and thereby cause major declines and deterioration for the entire European financial insurance sector. The reason for its non-inclusion is that we consider that a priori it would worsen the VaR estimates. The use of rolling parameters for the normal and empirical distributions is justified by the attempt to obtain an estimate of VaR that takes into consideration what has occurred in the most recent scenarios, thus providing more updated estimates of market volatility than those obtained by using the entire sample.

As shown in Table 2, the average return is positive in both cases, with a volatility (i.e., measured by the standard deviation of the returns) that has been reduced in the most recent period. Since both series are skewed to the right and show excess kurtosis, the hypothesis of normality of returns based on the Jarque Bera test is rejected (p-value < 0.05). The most extreme values occur between 1996 and 2012, especially in 2008.

Table 2
Summarized statistics of annual returns.

	1996–2012	2013 (January–October)
Average	4.98%	12.23%
Minimum	− 68.85%	− 11.10%
Maximum	71.12%	33.48%
Deviation	25.24%	10.60%
Percentile 0,5%	− 58.21%	− 9.50%
Jarque Bera (p-value)	0.00%	0.28%
Observations	4288	214

Table 3
p-Values associated with the different analyzed backtesting test.

	2013 (January–October)	
	Normal	Empirical
Observations	214	
Excesses	0	0
Z-test	0.1499	0.1499
Pearson's Q	0.0008	0.0008
Spearman's Rho	0.0000	0.0000
Kendall's Tau	0.0000	0.0000

Note. This table shows the p-values associated with the different backtesting tests of VaR in the out-of-sample period considered (i.e., 2013). Normal and empirical are the methods used to estimate VaR. Z-test, Pearson's Q, Spearman's Rho and Kendall's Tau are the test that we could estimate in the out-of-sample period.

4. Findings

First, we analyzed the function of failures of VaR at 99.5% in 2013, taking into account the normal and empirical model (see Table 3). No failures occur in this period under any of the models, so the model may be appropriate for setting capital requirements because of its conservatism. The expected number of failures given a VaR of 99.5% in one year is approximately 1.3; most of the tests cannot be used if no failures occur. The Z-test does not reject the adequate models (i.e., p-value of 0.1499), whereas according to Pearson's Q test the models are inadequate (i.e., p-value = 0.00), given the small number of failures, when the models are analyzed together at different VaR levels. Spearman's Rho and Kendall's Tau tests detected a relationship between the VaR estimate and actual returns (i.e., p-value < 0.05); the Tau coefficient for the empirical model (normal) is in 0.40 (0.30) and the coefficient Rho 0.43 (0.61). Since we cannot perform more tests for 2013, we must consider whether these models have appropriately estimated historical risk within the sample.

Fig. 1 allows observing what the dynamics of VaR is in relation to the distribution of historical losses within the sample, which is used to obtain a first approximation of the VaR model performance. Here, it can

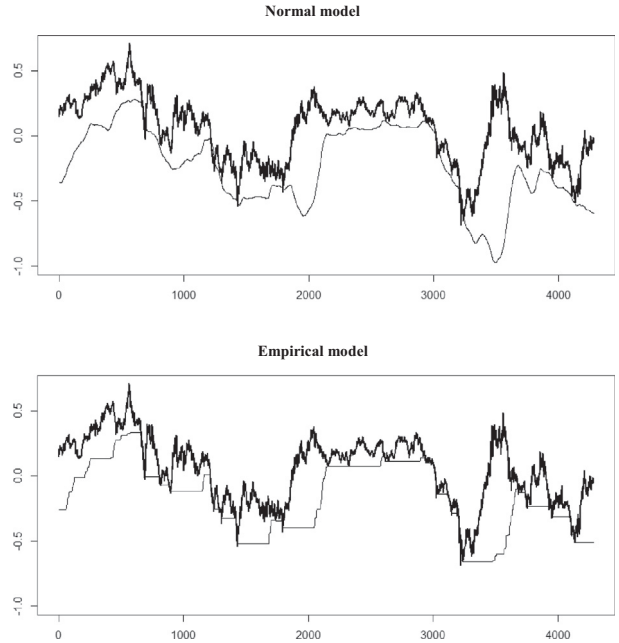

Fig. 1. Time evolution for the distribution of annual losses and of VaR (1996–2012).
Note. These figures show the historical pattern of losses and value-at-risk (VaR) for the normal and empirical model during the period from 1996 to 2012. Hits are the points where the loss is greater than VaR.

Table 4
p-Values associated with the different tests of backtesting.

	2008		1996–2012	
	Normal	Empirical	Normal	Empirical
Observations	253		4288	
Excesses	40	53	102	211
POF Kupiec	0.0000	0.0000	0.0000	0.0000
Christoffersen	0.0000	0.0000	0.0000	0.0000
Joint test	0.0000	0.0000	0.0000	0.0000
BCP k = 1	0.0000	0.0000	0.0000	0.0000
BCP k = 2	0.0000	0.0000	0.0000	0.0000
BCP k = 3	0.0000	0.0000	0.0000	0.0000
BCP k = 4	0.0000	0.0000	0.0000	0.0000
CAViaR β_0	0.0110	0.0002	0.0112	0.0000
CAViaR β_1	0.0000	0.0000	0.0000	0.0000
CAViaR β_2	0.3757	0.0665	0.0993	0.1473
Pearson's Q	0.0000	0.0000	0.0008	0.0000
Spearman's Rho	0.0000	0.0000	0.0000	0.0000
Kendall's Tau	0.0000	0.0000	0.0000	0.0000

Note. This table shows the *p*-values associated with the different backtesting tests of VaR in 2008 and in the overall in-sample period from 1996 to 2012. 2008 has been analyzed separately, because it was a stressed year affected by the financial crisis. Normal and empirical are the methods used to estimate VaR. POF, Christoffersen's (1998) test, Joint test, Berkowitz, Christoffersen, and Pelletier's (BCP) (2011) test, CAViaR, Pearson's Q, Spearman's Rho and Kendall's Tau are the test estimated in the in-sample period.

be seen that within the sample there have been different periods in which excesses have occurred, highlighting the period corresponding to 2008.

Table 4 shows the *p*-values associated with the backtesting tests explained in Section 2 for the normal and empirical model both for the entire analyzed period (from 1996 to 2012) and for 2008, which is considered a significant stress scenario. In 2008 (253 daily cumulative annual returns), a total of 40 excesses were produced in the standard model and 53 in the empirical model, a very high value considering the expected number; thus, the model is not appropriate based on Kupiec's (1995) test (i.e., *p*-value = 0.00). Many of these excesses are consecutive so the Christoffersen's (1998) test also rejects those models (i.e., *p*-value = 0.00). Given that both Kupiec's (1995) test and Christoffersen's (1998) test reject these models, the joint test reaches the same conclusion (i.e., *p*-value = 0.00). The BCP test detected autocorrelation for the four analyzed delays, which shows a strong time dependence structure among failures (i.e., *p*-value = 0.00). The CAViaR test finds that the β_0 parameter is statistically significant, indicating that the number of failures is not in line with the probability α, and the β_1 parameter, which again shows a positive relationship among failures (i.e., *p*-value = 0.00). Pearson's Q detects that the models are not adequate when jointly analyzed, by finding the VaR at the level of 99.5%, 97.5%, and 95% (i.e., *p*-value = 0.00). Spearman's Rho and Kendall's Tau tests detected a strong relationship between the carried out VaR estimate and real returns (i.e., *p*-value = 0.00), which was stronger than in the case of the empirical model. The Tau coefficient for the empirical model (normal) was at 0.93 (0.90) with a Rho coefficient 0.79 (0.70). The reader might think that these results may be due to the extreme of annual returns in 2008; it is therefore useful to analyze the results within a longer time period.

In the period from 1996 to 2012 (i.e., 4288 daily cumulative annual returns), a total of 102 excesses were produced in the normal model and 211 in the empirical model (see the last columns of Table 3), thus indicating again that the new models are inadequate based on Kupiec's (1995) test (i.e., *p*-value = 0.00). For Christoffersen's (1998) test, the BCP test for the four analyzed delays and the CAViaR test (the β_1 parameter takes the value 0.81 for the normal model and 0.65 for the empirical model) both find strong time-dependence relationship among failures. In addition, the parameter β_0 in the CAViaR test is once more statistically significant and positive, indicating that the number of failures is higher than expected. Pearson's Q test detects that the models are not suitable when analyzed together with different levels of confidence, as there are too many failures. Once more, Spearman's Rho and Kendall Tau tests detect a relationship between the VaR estimate and the real returns, which was even stronger than in the empirical model. The Tau model for the empirical coefficient (normal) is 0.53 (0.48) and the Rho coefficient is 0.68 (0.63). Thus, we can conclude that the number of failures in the models is excessive at different levels of confidence, that failures are auto-correlated and that there is a relationship (either positive or negative) between the VaR estimate and the returns.

The results show that even if a model seems to be conservative out-of-sample, it may nevertheless be inadequate judging by its historical performance. To the extent that the time horizon of investors in insurers is usually medium to long-term, we believe that a fundamental aspect to consider in the approval of an internal model is the model's suitability to represent historical scenarios. In addition, since the internal models should be periodically adapted to new developments, the adjustment within the sample with a certain time delay determines that the model has to be validated posteriori with the new data, which were previously outside the sample.

5. Conclusion

Solvency II allows using of internal models, approved by the regulator, based on certain requirements. Such models will have to overcome a validation test, whereby a set of tools (i.e., stress tests, scenario analysis, etc.) are grouped where backtesting is found. The objective of the backtesting is to demonstrate to the authorities that the resulting capital requirements are appropriate when comparing the expected results with the actual results. In this sense, Article 124 of Directive Solvency II and CEIOPS (2009) are inspired by an out-of-sample backtesting. Although we agree that theoretically this approach can be more appropriate, especially for banks, from a more practical point of view, we propose to extend in-sample backtesting for the insurance industry using a long-term horizon,. In this sense, we try to prove that a good model should not only try to measure future risk, but should also conform to the historical scenarios. This is partly reflected in Guideline 42 on the use of internal models (EIOPA, 2014), which establishes that an insurer must “validate the internal model under a wide range of circumstances that have occurred in the past or could potentially occur in the future.” Because a Solvency II backtesting should be performed at least annually, the option of using daily data with a temporary horizon of daily backtesting, as used in the banking sector, does not seem appropriate; thus, we employ annual returns.

This paper has proposed a practical methodology for backtesting, encompassing the adjustment of the two commonly used models, the normal and the empirical model, to calculate VaR. We used daily observations and accumulated annual returns using a rolling window of 252 days. The models fit their values based on the yields observed over the past year. We used time series of a Spanish equity index (i.e., IBEX35), where the sample has been divided into two periods and corresponding techniques: in-sample backtesting from 1992 to 2012 and out-of-sample backtesting from January to October 2013. The obtained results show that the models reach non-failure out-of-sample. When the results are analyzed in-sample, they are rejected on the basis of Kupiec's (1995) and Christoffersen's (1998) test, the Joint test, BCP test, a variant of the so-called CAViaR test, and a test of various levels (i.e., Pearson's Q test) and displayed an excessive number of failures. This leads us to conclude that a model should not only work well outside the sample, as is currently required in Solvency II, but must also be able to properly reflect the risk assumed in historical settings. Otherwise, a model could be judged suitable solely because the time period analyzed outside the sample is stable or slightly volatile. Since internal models should be regularly adapted to new developments, the adjustment within the sample, with some degree of time delay,

determines that the model should be validated retrospectively with new data, which initially were unavailable. In addition to recommending applying inside- and outside-of-sample backtesting, we also advise insurance companies to use different backtesting tests, since each of them measures different desired properties of the function of failures. Finally, as required by regulation, backtesting must be completed with other validation techniques such as stress testing or sensitivity analysis. Future research is necessary to pick the most suitable models and the inclusion of the dampener effect in determining capital requirements for equities (see Majri & De Lauzon, 2013). Another extension is to consider backtesting for Expected Shortfall (ES) or TailVaR (TVaR) instead of VaR, because ES is in general a better risk measure than VaR (Kratz, Lok, & McNeil, 2016). A limitation of our study is that to properly judge the outcome of a backtest when a model is backtested using a rolling window, the insurance company must keep in mind that the probability of finding an extreme event will increase and should therefore correct this increased probability (see Loois, 2015).

References

Berkowitz, J., Christoffersen, P., & Pelletier, D. (2011). Evaluating value-at-risk models with desk-level data. *Management Science, 57*, 2213–2227.

Campbell, S. (2005). *A review of backtesting and backtesting procedure.* Washington, D.C.: Federal Reserve Board. http://www.federalreserve.gov/pubs/feds/2005/200521/200521abs.html.

CEIOPS (2009). *CEIOPS' advice for Level 2 implementing measures on Solvency II: Articles 120 to 126, Tests and standards for internal model approval. CEIOPS-DOC-48/09*.

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

Christoffersen, P., & Diebold, F. (2000). How relevant is volatility forecasting for financial risk management? *The Review of Economics and Statistics, 82*, 12–22.

Christoffersen, P., Diebold, F., & Schuermann, T. (1998). Horizon problems and extreme events in financial risk management. *Economic Policy Review, 4*, 109–118.

Degiannakis, S., & Xekalaki, E. (2007). Assessing the performance of a prediction error criterion model selection algorithm in the context of ARCH models. *Applied Financial Economics, 17*, 149–171.

Dowd, K., Blake, D., & Cairns, A. (2004). Long-term value at risk. *The Journal of Risk Finance, 5*, 52–57.

EIOPA (2014). *EIOPA-BoS-14/180: Guidelines on the use of internal models*.

Eling, M., & Pankoke, D. (2014). Basis risk, procyclicality, and systemic risk in the solvency II equity risk module. *Journal of Insurance Regulation, 33*, 1–39.

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

Escanciano, J. C., & Pei, P. (2012). Pitfalls in backtesting historical simulation VaR models. *Journal of Banking & Finance, 36*, 2233–2244.

Evers, C., & Rohde, J. (2014). Model risk in backtesting risk measures (working paper). http://diskussionspapiere.wiwi.uni-hannover.de/pdf_bib/dp-529.pdf.

Fedor, M. (2007). Economic capital versus regulatory capital for market risk in banking and insurance sectors: Basel II experience and the challenge for Solvency II. http://www.actuaries.org/AFIR/Colloquia/Stockholm/xFedor.pdf.

Kratz, M., Lok, Y. H., & McNeil, A. J. (2016). Multinomial VaR backtests: A simple implicit approach to backtesting expected shortfall (working paper). https://arxiv.org/abs/1611.04851.

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

Lehikoinen, K. (2007). *Development of systematic backtesting processes of value-at-risk* (Master's Thesis)Helsinki University of Technologyhttps://pdfs.semanticscholar.org/0c6c/64126a2d5108e47c8823e995f381d3112753.pdf.

Loois, M. (2015). Backtesting solvency II value-at-risk models using a rolling horizon. *Journal of Risk Model Validation, 9*, 13–31.

Majri, M., & De Lauzon, F. X. (2013). An effective equity model allowing long term investments within the framework of Solvency II. http://hal.archives-ouvertes.fr/docs/00/84/78/87/PDF/MAJRI_20130515.pdf.

Moody's Analytics (2013). Validation of risk factor modelling in 1-year VaR capital assessments. http://www.moodysanalytics.com/~/media/whitepaper/2013/2013-10-04-Validation-of-risk-factor-modelling-in-1-year-VaR-capital.pdf.

Otero, L., Durán, P., Fernández, S., & Vivel, M. (2012). Estimating insurer's capital requirements through Markov switching models in the solvency II framework. *International Journal of Finance and Economics, 86*, 20–38.

Panning, W. H. (1999). The strategic uses of value at risk: Long-term capital management for property/casualty insurers. *North American Actuarial Journal, 3*, 84–105.

Piontek, K. (2009). The analysis of power for some chosen VaR backtesting procedures: Simulation approach. In A. Fink, B. Lausen, W. Seidel, & A. Ultsch (Eds.). *Advances in data analysis, data handling and business intelligence* (pp. 481–490). Berlin, Heidelberg: Springer-Verlag.

So, M. K. P., & Yu, P. L. H. (2006). Empirical analysis of GARCH models in value at risk estimation. *Journal of International Financial Markets, Institutions & Money, 16*, 180–197.

Virdi, N. K. (2011). A review of backtesting methods for evaluating value-at-risk. *International Review of Business Research Papers, 7*, 14–24.