New Methods in Backtesting

Marcus Haas

Financial Engineering
Research center caesar
Friedensplatz 16, D-53111 Bonn

February 5, 2001

Abstract

To evaluate the goodness of a VaR-model, banks as well as regulators use backtesting to confirm their judgments. Banks find it of special interest to know about the outcome of a backtest before it is conducted by the regulators, thus detecting problems in the VaR-process in time and preventing fines. This paper summarizes existing methods, offers improvements to these methods and tries to find an optimal backtesting strategy.

Introduction

When it comes to backtesting, banks are often interested in simply fulfilling the regulatory requirements. Therefore, conservative methods are pursued, which might be more "expensive" than strategies which are close to the market and are fined from the regulatory side. Backtesting offers a variety of methods, which extend way beyond the boundaries of the Basle traffic light. Specific backtests can identify possible flaws of VaR-models. Backtesting can also tell us something about the investment strategy of a portfolio manager and about the dynamics of a portfolio. In chapter I, we will take a closer look at the existing regulatory requirements. Chapter II introduces some current backtesting methods. Chapter III presents new improved backtesting methods. Chapter IV shows some results and interpretations and chapter V summarizes and gives directions for optimal backtesting.

1 Regulatory Requirements

Since 1996, the Basle Committee on Banking Supervision instructs banks (and other companies, like insurances) to develop their own risk models to evaluate their portfolio risk. Central to this models is the so-called "Value at Risk" (VaR), which describes the maximum portfolio loss over a given time horizon - usually one day - for a given probability. A 99%-VaR of $100.000 for example means that the probability that the portfolio loss on the next day will exceed $100.000 is less or equal one percent. If the VaR-number is undershot by the portfolio movement, we speak of an "exception". According to the number of exceptions, a scaling factor is identified which, multiplied by the VaR-number, determines the amount of capital to be held by the bank or company. If the VaR is calculated at the 99% level, we can expect an exception every 100 days. This assumption is the base of the Basle traffic light, the official backtesting model. If the exceptions are independent, the probability for an exception is 100-(VaR-level)% per day (in our example 1%). The Basle traffic light demands a time series of 250 days (reflecting one year), thus we can expect between 2 and 3 exceptions over this time period, given the 99%-VaR. Mathematically speaking, we model the event of a "VaR-exception" as a binomially distributed B(250;0,01) rv. It's expected value is 2,5. Presuming that the expected value will not be matched with probability 1 and that the result will also be falsified by wrong assumptions like the independence of the exceptions, we already have a justification for the Basle traffic light zones. It demands that the multiplication factor is raised, if we experience more than five exceptions. The following table describes the scaling factors of the Basle traffic light.

Exceptions	< 5	5	6	7	8	9	> 9
Scaling factor	3	3.4	3.5	3.65	3.75	3.85	4

It is apparent, that the Basle traffice light approach cannot be used to evaluate the goodness of a VaR-model. It neither considers the measure of the exceedance, nor its position. In practice, there are some tests that treat these subjects. We will briefly present them in the next chapter.

2 Existing methods

We will now have a look at existing backtesting methods. Some of them, like the Kupiec tests are based on the same binomiality assumption of the exceptions as the Basle approach. Others take the measure of the exceedance into account or judge the VaR-model as a whole.

2.1 Kupiec's POF-Test (Proportion of Failures)

Both of the Kupiec-tests are so-called "Likelihood-Ratio-Tests". The null hypothesis for these tests is, that the empirically determined probability matches the given probability. If we stay with our example and assume, that the VaR has been calculated at the 99% level, we will test for $H_0 : p = \tilde{p} = x/n = 0.01$, where x represents the number of exceptions and n represents the number of backtesting points. We can now use a Likelihood-Ratio-Test to check for this assumption. The corresponding LR-statistic is defined as

$$LR_{\text{POF}} = -2ln\left(\frac{p^x(1-p)^{n-x}}{\tilde{p}^x(1-\tilde{p})^{n-x}}\right).$$

It is asymptotically χ^2 distributed with one degree of freedom. If the value of the LR-statistic exceeds a critical value, e.g. the 95%-quantile of the χ^2_1-distribution, we will decline the H_0-Hypothesis, otherwise we will accept it. Christoffersen continues this thought and introduces a mixed test which combines the POF-Test with a test for independence. He thus receives a test statistic which is χ^2 distributed with 2 degrees of freedom. However, his test for independence is too weak to deliver feasable results. In chapter III, we will introduce an improved test for independence and coverage.

2.2 Kupiec's TUFF-Test (Time until First Failure)

This test is based on similar assumptions as the POF-Test. If we take the exceptions to be binomially distributed, then the probability of an exception is again the inverse probability of the VaR confidence-level, in our case 1%. Thus we will also expect an exception every 100 days. We can now use this to create an LR-Test which measures the time until the first exception. The null hypothesis is set to $H_0 = p = \hat{p} = 1/\nu = 0.01$ where ν is the time until the first exception in our sample. The corresponding LR-statistic is defined as

$$LR_{\text{TUFF}} = -2ln\left(\frac{p(1-p)^{\nu-1}}{\hat{p}(1-\hat{p})^{\nu-1}}\right).$$

It is also asymptotically χ^2-distributed with one degree of freedom and we can confirm the H_0-hypothesis in the same way as before.

2.3 Point estimator for p

Under the same assumptions as before with the two Kupiec-tests, we can directly derive an estimator $\hat{p}$ from the number of exceptions in our sample and compare it to the given value p. The test results will not differ much from the Kupiec POF-test, however we can now judge the degree of error of our model parameter and gain knowledge about the true value of p. Thus we are using the Maximum-Likelihood-Estimator

$$\hat{p} = \frac{x}{n}$$

with variance

$$V(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n}.$$

If we now consider the interval $\left[\hat{p} - \sqrt{V(\hat{p})}, \hat{p} + \sqrt{V(\hat{p})}\right]$, we have derived a confidence interval for the estimator of p. If p lies within this interval, we can consider the model to be good, if it lies outside, we can judge which confidence level our model would have rendered.

2.4 Lopez' Magnitude Loss Function

Contrary to the other tests, Lopez suggests a testing method which also incorporates the magnitude of the exceptions in addition to their number. He introduces the variable C, which takes the sum of the number of exceptions and their squared distance from the corresponding VaR as a value:

$$C_i = \begin{cases} 1 + (x_i - \hat{v}_i)^2, & x_i \leq \hat{v}_i \\ 0, & x_i > \hat{v}_i \end{cases} \quad i \in \{1, 2, \ldots, n\} \ .$$

Here, $\hat{v}_i$ represents the VaR-forecast for the ith day and x_i reflects the portfolio movement on day i. We can now compare the value of C to a benchmark value. To calculate this benchmark, we simulate many (e.g.1000) P&L time-series from our model and calculate the corresponding C-value. Then we take a high quantile from these C-values (e.g. 80%) and check if our actual value lies below or above this quantile value. If it lies below, we accept the model, if it lies above, we decline it.

2.5 Test from Crnkovic and Drachman

Contrary to other methods, the CD-test not only evaluates the exceptions but looks at the entire VaR-model instead. Therefore, empirical percentiles are calculated for every portfolio movement:

$$p_i = F(x_i),$$

where F represents the modelling distribution. If the model is well calibrated we can expect every percentile value in $[0, 1]$ with the same probability. Thus, if we look at a series of values of p_i we should not be able to tell them apart from a series of realizations of uniformly distributed rvs. Thus our testing hypothesis is formulated as

$$P_i \sim R(0,1) \text{ iid.}$$

We may varify this assumption through a number of tests. Crnkovic and Drachman suggest a Q-test for the distributional assumptions (this test compares the maximum distances to the uniform distribution with a benchmark) and a BDS-test for independence.
Introducing a "worry-function", we can furthermore focus the tails of the distribution. Crnkovic and Drachman suggest using $f(t) = 0.5ln(t(1-t))$. The corresponding critical value for the Q-test will then have to be computed with a Monte-Carlo simulation.

3 Improved Backtesting

To start the discussion about new and improved backtesting methods, let us have a look at the disadvantages of the existing methods. Both of the Kupiec tests are by themselves not powerful enough to correctly predict errors in the VaR-model. Kupiec has already mentioned this problem in his own paper. The problem of the Lopez test is, that sudden and very extreme events, that cannot be coped with in any model will create very high values of C which will decline every model. If one is interested in an automization of backtesting, the Lopez-Test is probably the least suited. The CD-Test is the most flexible of the tests mentioned before, however it also has some very strong requirements, such as knowledge of the distribution function underlying the model at any given point in time. If the VaR is calculated with a historical simulation for example, this fact is not given. In these cases, the percentiles can only be derived numerically, which is computationally challenging as well as very unprecise. Furthermore, the tests for a uniform distribution are quite weak and according to Crnkovic and Drachman, sample sizes of at least 1000 points should be used to make reliable judgments. To avoid these problems we will now reduce the CD-method and deploy a scaled CD-method.

3.1 Scaled CD-Method

Given the distribution function for each day, we can directly calculate the percentiles for the CD-method. If that is not the case however, this turns out to be quite challenging. Thus, instead of the exact percentiles, we calculate quantiles (using the VaR-model). We do this for a certain number of quantiles and thereby split the real axis into a number of intervals which are all equally likely to be hit by the portfolio movement. Let us illustrate this with an example. Assuming we divide the real axis into twenty intervals I_1 to I_{20}. To do so, we use the VaR-model to calculate every quantile with 5 percent distance, starting with the 5%-quantile (then the 10%, 15% and so on). We derive the following partition:

$$I_i = \left[q_{\frac{i-1}{20}}, q_{\frac{i}{20}}\right), i = 2, ..., 19$$

and

$$I_1 = (-\infty, q_{\frac{1}{20}}), \;\; I_{20} = \left[q_{\frac{19}{20}}, \infty\right)$$

The actual portfolio movement will now hit one of these twenty intervals. Similar to the actual CD-method, the probability for the event “Portfolio movement hits interval i” is the same for all intervals. If we could calculate the percentiles directly (using the distribution function of the VaR-model), we can now create a histogram of these percentiles and also derive the interval hits. Figure 1 shows the plot of a scaled CD-procedure with 500 backtesting points and 20 intervals using the Exxon-share quotes from to .

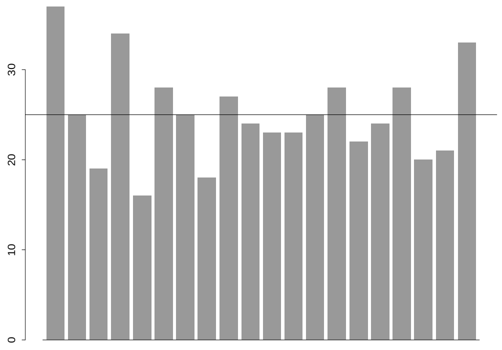


Figure 1: Scaled CD-method of a historical simulation of the Exxon-stock

From this plot, we can already learn about several aspects of the portfolio. For example, the historical simulation might not be an adequate model in this case. The first interval, reflecting the 95%-VaR has significantly more hits than expected. The same can be said about the last interval. From this plot, we can also judge, that the volatility is increasing during the observed time span. Comparing the plot of the sorted percentiles to a scaled CD-plot, we can observe the significance of this method. Figure 2, which is a numerically created plot of the sorted percentiles for the same Var-simulation as before can barely be told apart from a straight line, representing the uniform distribution.

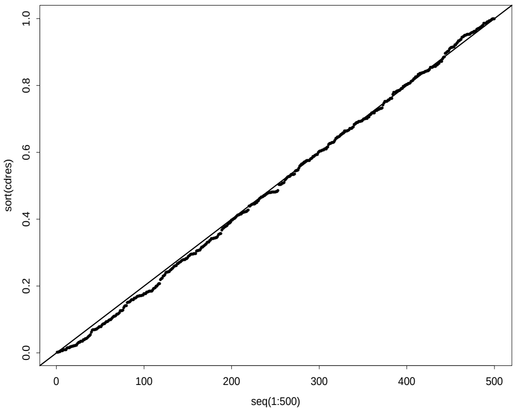


Figure 2: Sorted percentiles of a historical VaR-simulation of the Exxon-stock

The main advantage over the ordinary CD-method is, that we no longer have to judge a continous distribution, but a discrete one instead. More precisely, we are now testing for a multinomial distribution $M(n, p)$, where n is the number of backtesting points, and p is the vector of hitting probabilities for each of the intervals. In our example, a vector which contains the value $\frac{1}{20}$ twenty times. We can now run a number of tests to verify this distributional assumption. The most suitable test is a χ^2-distance test. We will be testing the null hypothesis

$$H_0 : (p_1, \ldots, p_r) = (p_1^0, \ldots, p_r^0)$$

against the alternative

$$H_1 : (p_1, \ldots, p_r) \neq (p_1^0, \ldots, p_r^0),$$

where r is the number of intervals. The corresponding test statistic is

$$Q(Y_1,\dots,Y_r;p_1^0,\dots,p_r^0) = \sum_{i=1}^{r} \frac{(Y_i - np_i^0)^2}{np_i^0} = \sum_{i=1}^{r} \frac{Y_i^2}{np_i^0} - n,$$

where Y is the random variable corresponding to y. If we assume, that all intervals are hit with equal probability, the null hypothesis comes down to

$$H_0 : (p_1,\dots,p_r) = \left(\frac{1}{r},\dots,\frac{1}{r}\right)$$

and the test statistic to

$$Q\left(Y_1,\dots,Y_r;\frac{1}{r},\dots,\frac{1}{r}\right) = -n + \frac{r}{n}\sum_{i=1}^{r} Y_i^2.$$

Q is asymptotically χ^2-distributed with $r-1$ degrees of freedom. If the value of Q for a realization y of Y exceeds a certain critical value, e.g. the 95%-quantile of the χ^2_{r-1}-distribution, we decline the model, otherwise we accept it. Much more meaningful is of course the graphical analysis of the interval hits, as we have already mentioned before.

Similar to the ordinary CD-test, we can introduce a weighing scheme in order to focus on the outer quantiles. To do so, we decrease the size of the outer intervals and receive a new nesting which we can now run a χ^2-test for. We have to pay attention though to the fact, that the parameters p_1 to p_r are now different from each other and correspond to the interval length we have associated to them.

Let us finish this chapter on the scaled CD-method with an example. We are looking at stock quotes for the BMW-share and calculate the corresponding VaR with a VCV-approach for 500 backtesting points with a history of 500 points. Other backtests, like the Kupiec POF-test and a graphical analysis suggest, that the VCV-approach might render bad results for the given time horizon. We will try to confirm this result with the scaled CD-method. We will begin by looking at a plot of the interval hits for 20 equalsized intervals.

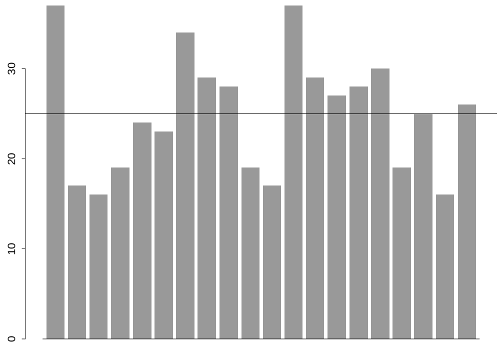


Figure 3: Scaled CD-method of a VCV-VaR-simulation of the BMW stock

This picture already suggests, that the VCV-model is completely unfit for the VaR-calculation in this case. The high bar representing the 5%-quantile and thus the exceedances of the 95%-VaR suggests a gross underestimation of the risk. At the same time, this mass seems to be missing in the middle, leading us to believe that in more quite times, the VaR was too conservative. If we look at the value of the χ^2-test-statistic, we receive 34.08, which just barely exceeds the critical value of 30.14, but we still have a decline. We now use a weighted test by choosing the intervals such that their length is half that of the adjacent inner interval. The two innermost intervals thus have the length of $\frac{1}{4}$ in probability, the two intervals next to it $\frac{1}{8}$ and so on, with the outermost intervals taking the value of $\frac{1}{64}$ in probability. This yields us 12 intervals and the corresponding χ^2-statistic now takes a value of 27.19, exceeding the critical value of 19.68 by quite a margin.

In summary, we can say that the scaled CD-method not only renders new graphical analyses, but also allows statements about the entire VaR-model, with or without focussing on the tails, in a simple fashion.

3.2 Mixed Kupiec-Test

Secondly, we are suggesting a mixed test which picks up the ideas of Christoffersen, but uses more powerful tests to render better results. Let us recapitulate the results of the Kupiec TUFF-test. Its test statistic was asymptotically χ^2-distributed with one degree of freedom and measures the time until the first exception. In the same way, we can measure the time between two exceptions, resulting in the following test statistic:

$$LR_i = -2ln\left(\frac{p(1-p)^{\nu_i - 1}}{\hat{p}(1-\hat{p})^{\nu_i - 1}}\right).$$

We use the same notation as before, except that ν_i is now the time between exception i and exception $i-1$. If our model is optimal, then we can once again expect an exception to occur every 100 days. We can thus construct a "Time between failures"-test for every exception and an additional TUFF-test for the first one. As a result, we will receive n test statistics, where n is the number of exceptions. Since our null hypothesis is that the exceptions are independent from each other, the test statistics are independent as well and we can sum them up. The χ^2-distribution is additive as well, so we can also add the critical values. As a result, we receive a test for independence, where we test the null hypothesis H_0 "The exceptions are independent from each other". If we have a total of n exception, the corresponding test statistic is as follows:

$$LR_{ind} = \sum_{i=2}^{n}\left(-2ln\left(\frac{p(1-p)^{\nu_i - 1}}{\hat{p}(1-\hat{p})^{\nu_i - 1}}\right)\right) - 2ln\left(\frac{p(1-p)^{\nu - 1}}{\hat{p}(1-\hat{p})^{\nu - 1}}\right)$$

Asymptotically, this is a sum of n rvs, where each is χ^2_1-distributed, which results in a χ^2_n-distributed rv. The test statistic from the Kupiec POF-test is also independent from all the test statistics mentioned above, so we can add it as well. As a result, we receive a new mixed test for independence and coverage, which has a test statistic of

$$LR_{mix} = LR_{ind} + LR_{pof} \sim \chi^2_{n+1}$$

Just like we would do with the other Kupiec tests, we can now compare the value of LR_{mix} to the 95%-quantile of the χ^2_{n+1}-distribution. If it is lower, we will accept the model, if it is higher, we decline it. Detailed examples for this can be found in the next chapter.

4 Backtesting results

We want to observe how well the previously introduced tests are suited to judge the quality of a VaR-model. As a standard gauge, we use a graphical analysis of the exceptions as well as the results of a point estimation and the Basle traffic light. Point of examination are a VCV-approach as well as a historical simulation with 250, 500 and 1000 points and a history of 300, 500 and 1000 points respectively. The data used for modelling are stock quotes from BMW, Exxon and the German DAX index, which were chosen at random.

BMW - VCV-Approach

Points	250	500	1000
History			
300	(G)+ + +	(Y-)- - -	(-)- - -
500	(G)+ + +	(R)- - -	(-)- - -
1000	(G)+ + +	(R)- - -	(-)- - -

BMW - hist.Simulation

Points	250	500	1000
History			
300	(G)+ + +	(Y)- + +	(-)- - -
500	(G)+ + +	(Y-)- + +	(-)- - -
1000	(G)+ - -	(R)- - -	(-)- - -

EXXON - VCV-Approach

Points	250	500	1000
History			
300	(G)- + +	(G)+ + +	(+)+ + +
500	(G)+ + +	(Y-)- + +	(+)+ - -
1000	(Y)+ + -	(Y-)- - -	(-)- - -

EXXON - hist.Simulation

Points	250	500	1000
History			
300	(G)+ + +	(Y+)+ + +	(-)+ + -
500	(Y+)+ + +	(Y-)- + +	(-)- - -
1000	(Y-)- + -	(Y-)- - -	(-)- - -

DAX - VCV-Approach

Points	250	500	1000
History			
300	(G)+ - -	(Y)- + +	(-)- - -
500	(G)+ + +	(Y-)- + -	(-)- - -
1000	(Y+)+ + +	(R)- - -	(-)- - -

DAX - hist.Simulation

Points	250	500	1000
History			
300	(G)+ + -	(Y)- + +	(-)- - -
500	(G)+ + +	(Y)- - -	(-)- - -
1000	(G)+ + +	(Y-)- - -	(-)- - -

In brackets are the results of the Basle traffic light. If the result is “Yellow” or the Basle scheme cannot be utilized to make a judgment (at 1000 points), a “+” or “–” sign indicates if the other tests suggest an acceptance or a declination. The tests utilizied for this are the point estimator and a graphical analysis, as mentioned above. The three signs outside the bracket indicate from left to right the results of the mixed Kupiec test, scaled CD-test with 20 equal-sived intervals and the weighted scaled-CD test with 12 intervals as explained before.

If we have a look at the results, we can immediately see, that the “red” models were also declined by all three of the other tests. The “green” models were also mostly identified. If the Kupiec-test declines the green model, it is because dependencies are not compensated in the Basle scheme. Declinations in the CD-test are caused by too conservative VaR-models. The most interesting events however occur, when the traffic light is “yellow” and the model cannot be quantified as good or bad from a regulatory point of view. If we consider the results of the other backtests, then those are mostly confirmed by the Mixed Kupiec test. The scaled CD-test can then tell us why the judgment was indecisive before. If it is negative, then the model is badly calibrated altogether. If the weighted scaled CD-test is negative, this might have several reasons, like a mismodelling of the high gains or the high quantiles altogether. A graphical analysis usually answers these questions right away. To finish off this chapter, let us have a look at two explicit examples, beginning with the result for the historical simulation of BMW with 500 backtesting points and 300 history.The result “(Y) - + +” is indecisive. From the result we are assuming a cluster of exceptions and an otherwise well-calibrated model. Looking at figure 4, this judgment is only partially correct.

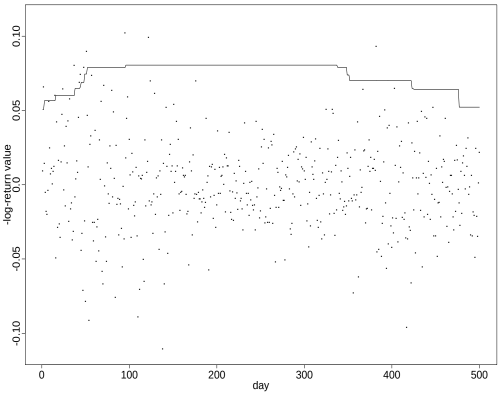


Figure 4: 500-day-VaR and P&L for BMW with 300 points history

Apparently, the strongly varying volatility is recorded by the VaR-model too slowly, thus creating a phase with too many exceptions as well as a phase with a too conservative VaR. Our first impression is therefore only partially confirmed, because even though the CD-test was accepted, the model is not well calibrated - a problem which occured with the regular CD-test as well.

Secondly, we want to have a look at the results of a historical simulation for the Exxon stock with 1000 backtesting points and 300 history. They are "(-) + + -". From this, we conclude that the model is well calibrated, but that the VaR is generally underestimated. This impression is confirmed by figure 5.

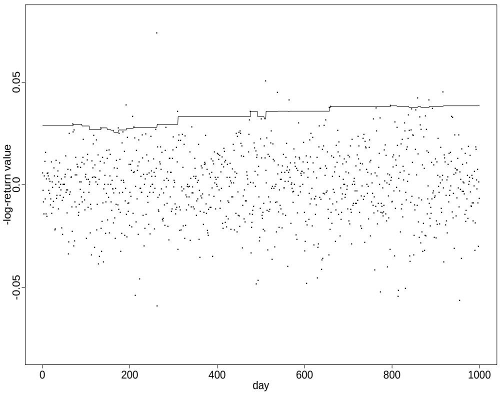
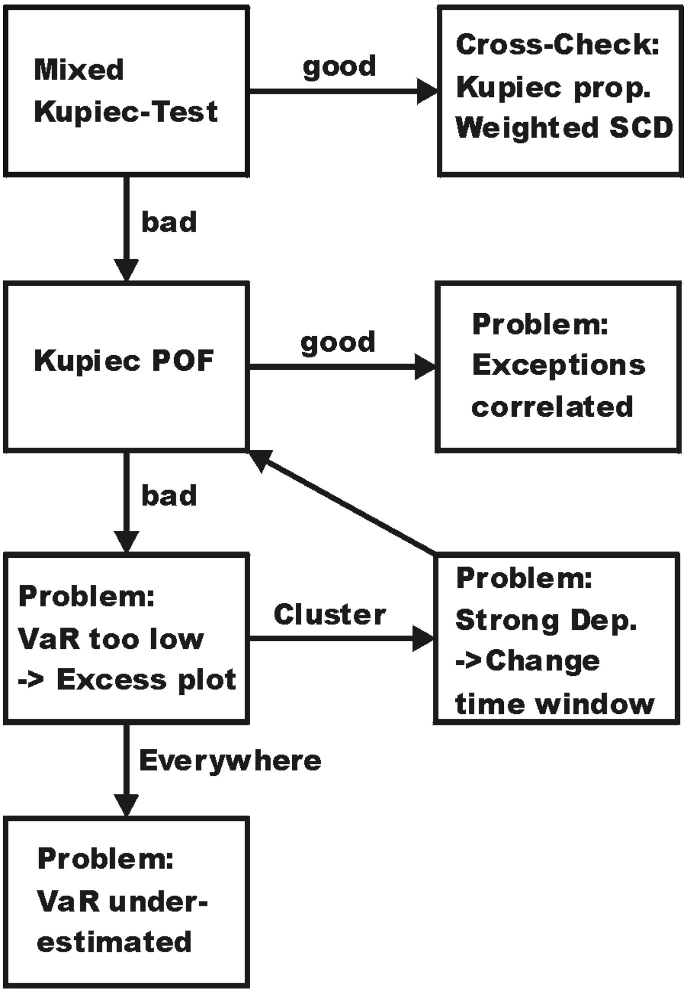


Figure 5: 1000-day-VaR and P&L for EXXON with 300 points history

In general, we can say that good or bad results in the mixed Kupiec or scaled-CD test can also be confirmed through other backtests. When the results differ however, another test should be included for confirmation.

5 Summary and outlook

Finally, we would like to use the newly introduced methods to set up strategies for optimal backtesting. We can definitely say, that a single backtest can never be enough to judge the quality of a VaR-model. Good results in one test should thus always be confirmed by another test. We will clarify this in the following diagram:


Mixed Kupiec-Test
good
Cross-Check: Kupiec prop. Weighted SCD
bad
Kupiec POF
good
Problem: Exceptions correlated
bad
Problem: VaR too low -> Excess plot
Cluster
Problem: Strong Dep. ->Change time window
Everywhere
Problem: VaR under-estimated

From the two newly introduced tests, the mixed Kupiec-test appears to be the strongest, since it can identify both problems with dependencies and the number of exceptions. The scaled CD-test allows for judgment about the entire VaR-model, but can easily be fooled by dependencies within the data.

Combining both tests, we have the possibility of evaluating the model itself as well as assessing the number of exceptions relevant to the regulators. An optimal backtest cannot be achieved with these methods either and will always require an additional graphical analysis.

Acknowledgements

The author would like to thank that Westdeutsche Landesbank, especially Frank Bock and Claudia Keseling for their support and incentives. The results of this paper and the corresponding software tool are currently in use at the Westdeutsche Landesbank.

References

[Basel96] Basler Ausschuss für Bankenaufsicht *Änderung der Eigenkapitalvereinbarung zur Einbeziehung der Marktrisiken*, Bank für Internationalen Zahlungsausgleich, Basle 1996

[CGG97] Cassidy, C. and Gizycki, M. *Measuring Traded Market Risk: Value-at-Risk and Backtesting Techniques*, Reserve Bank of Australia, 1997

[Christoffersen98] Christoffersen, P. *Evaluating Interval Forecasts*, McGill University, Canada, published in: International Economic Review, Vol.39 No.4, November 1998

[CD96] Crnkovic, C. and Drachman, J. *Quality Control*, RISK, 9, 1996

[KS99] Krämer, M and Schmidt, H. *Konsistentes Backtesting - ein Vergleich verschiedener Verfahren*, Handbuch Bankenaufsicht und Interne Risikomodelle, Schöffer/Poeschel, 1999

[Kupiec95] Kupiec, N.H. *Techniques for Verifying the Accuracy of Risk Measurement Models*, Journal of Derivatives, Vol.3, Nr.2, 1995

[Lopez98] Lopez, J.A. *Methods for Evaluating Value-at-Risk Estimates* Federal Reserve Bank of New York, 1998

[OS00] Overbeck, L. and Stahl, G. *Backtesting: Allgemeine Theorie, Praxis und Perspektiven*, BAKred, Bonn and Deutsche Bank, Preprint 2000

[OZ99] Overbeck, L. and Zakrzewski, O. *Der Q-Test von J.P.Morgan - Inhalt, Aussagekraft und Grenzen*, Handbuch Bankenaufsicht und Interne Risikomodelle, Schöffer/Poeschel, 1999

[SSM97] Schröder, F. and Schulte-Mattler, H. *CD-Verfahren als Alternative zum Basler Backtesting*, published in: Die Bank, 7/97