Backtesting

FNCE 5321

Hang Bai

Overview

- The objective is to consider the ex ante risk measure forecasts from the model and compare them with the ex post realized portfolio return
- The backtest procedures can be seen as a final diagnostic check on the aggregate risk model, thus complementing the other various specific diagnostics

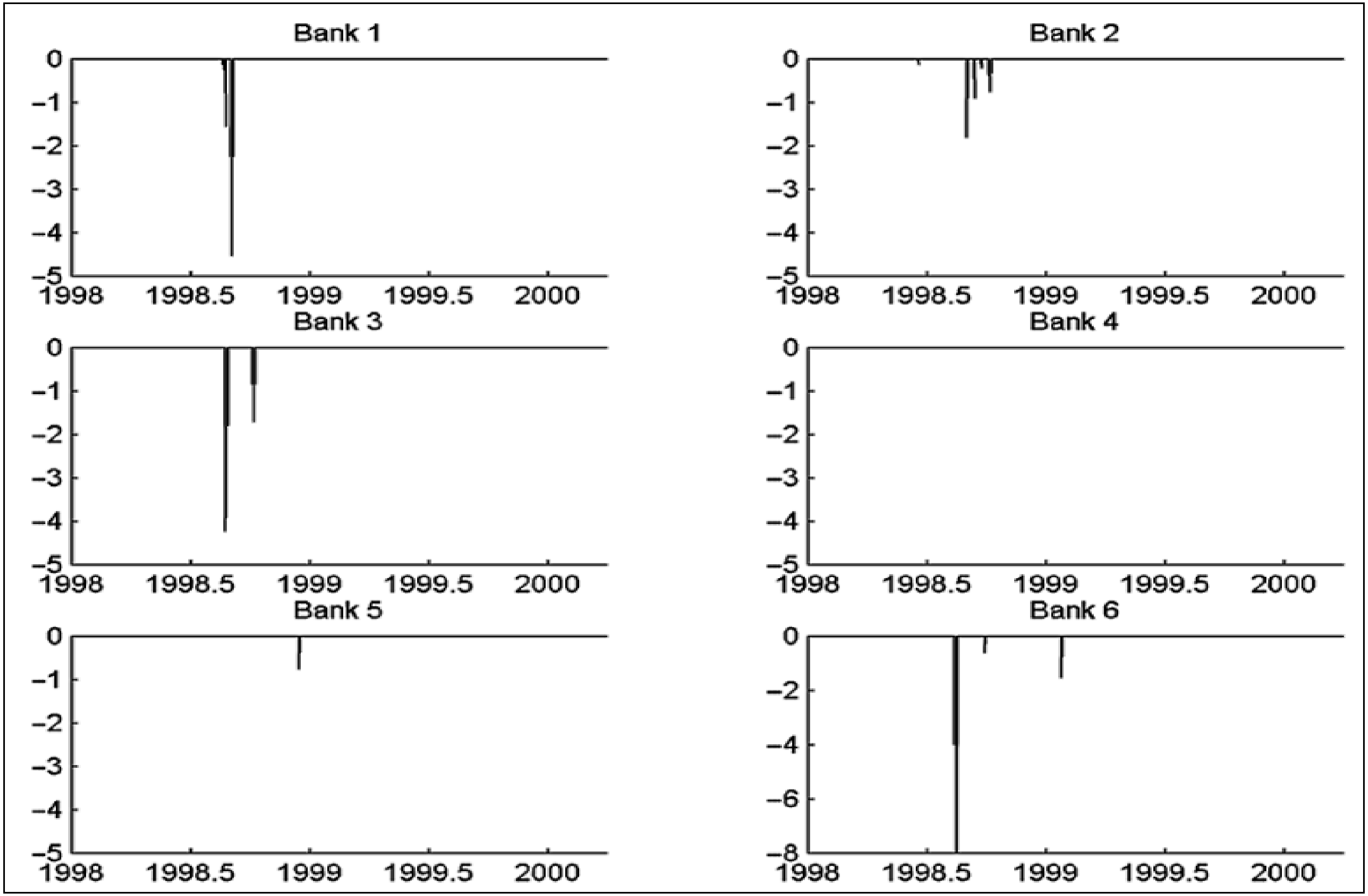
Overview

- We establish procedures for backtesting *VaR*s
- We start by introducing a simple unconditional test for the average probability of a *VaR* violation
- We then test the independence of the *VaR* violations
- Finally, combine unconditional test and independence test in a test of correct conditional *VaR* coverage

Overview

- Figure 13.1 shows the performance of some real-life *VaR*s
- Figure shows the exceedances of the *VaR* in six large U.S. commercial banks during the January 1998 to March 2001 period

Figure 13.1: Value-at-Risk Exceedences From Six Major Commercial Banks



Overview

- Whenever the realized portfolio return is worse than the *VaR*, the difference between the two is shown
- Whenever the return is better, zero is shown
- The difference is divided by the standard deviation of the portfolio across the period
- The return is daily, and the *VaR* is reported for a 1% coverage rate.
- To be exact, we plot the time series of

$$Min\left\{R_{PF,t+1}-\left(-VaR_{t+1}^{.01}\right),0\right\}/\sigma_{PF,t+1}$$

Overview

- Bank 4 has no violations at all, and in general the banks have fewer violations than expected
- Thus, the banks on average report a *VaR* that is higher than it should be
- This could either be due to the banks deliberately wanting to be cautious or the *VaR* systems being biased
- Another culprit is that the returns reported by the banks contain nontrading-related profits, which increase the average return without substantially increasing portfolio risk

Overview

- More important, notice the clustering of *VaR* violations
- The violations for each of Banks 1, 2, 3, 5, and 6 fall within a very short time span and often on adjacent days
- This clustering of *VaR* violations is a serious sign of risk model misspecification
- These banks are most likely relying on a technique such as Historical Simulation (HS), which is very slow at updating the *VaR* when market volatility increases
- This issue was discussed in the context of the 1987 stock market crash in Chapter 2

Overview

- Notice also how the *VaR* violations tend to be clustered across banks
- Many violations appear to be related to the Russia default and Long Term Capital Management bailout in the fall of 1998
- The clustering of violations across banks is important from a regulator perspective because it raises the possibility of a countrywide banking crisis
- Motivated by the sobering evidence of misspecification in existing commercial bank *VaR*s, we now introduce a set of statistical techniques for backtesting risk management models

Backtesting *VaR*s

- Recall that a VaR^p_{t+1} measure promises that the actual return will only be worse than the VaR^p_{t+1} forecast p . 100% of the time

- If we observe a time series of past ex ante *VaR* forecasts and past ex post returns, we can define the “hit sequence” of *VaR* violations as

$$I_{t+1} = \begin{cases} 1, if \ R_{PF,t+1} < -VaR^p_{t+1} \\ 0, if \ R_{PF,t+1} \geq -VaR^p_{t+1} \end{cases}$$

Backtesting *VaR*s

- The hit sequence returns a 1 on day *t*+1 if the loss on that day was larger than the *VaR* number predicted in advance for that day

- If the *VaR* was not violated, then the hit sequence returns a 0

- When backtesting the risk model, we construct a $\{I_{t+1}\}_{t=1}^{T}$e . across *T* days indicating when the past violations occurred

The Null Hypothesis

- If we are using the perfect *VaR* model, then given all the information available to us at the time the *VaR* forecast is made, we should not be able to predict whether the *VaR* will be violated

- Our forecast of the probability of a *VaR* violation should be simply *p* every day

- If we could predict the *VaR* violations, then that information could be used to construct a better risk model

The Null Hypothesis

- The hit sequence of violations should be completely unpredictable and therefore distributed independently over time as a Bernoulli variable that takes the value 1 with probability *p* and the value 0 with probability (1-p)
- We write:

$$H_0 : I_{t+1} \sim i.i.d.\ \text{Bernoulli}(p)$$

- If *p* is 1/2, then the i.i.d. Bernoulli distribution describes the distribution of getting a “head” when tossing a fair coin.
- The Bernoulli distribution function is written

$$f(I_{t+1}; p) = (1-p)^{1-I_{t+1}} p^{I_{t+1}}$$

The Null Hypothesis

- When backtesting risk models, *p* will not be 1/2 but instead on the order of 0.01 or 0.05 depending on the coverage rate of the *VaR*

- The hit sequence from a correctly specified risk model should thus look like a sequence of random tosses of a coin, which comes up heads 1% or 5% of the time depending on the *VaR* coverage rate

Unconditional Coverage Testing

- We first want to test if the fraction of violations obtained for a particular risk model, call it π, is significantly different from the promised fraction, *p*
- We call this the unconditional coverage hypothesis
- To test it, we write the likelihood of an i.i.d. Bernoulli(π) hit sequence

$$L(\pi) = \prod_{t=1}^{T} (1-\pi)^{1-I_{t+1}} \pi^{I_{t+1}} = (1-\pi)^{T_0} \pi^{T_1}$$

- where *T*0 and *T*1 are number of 0s and 1s in sample
- We can easily estimate π from $\hat{\pi} = T_1/T$; that is, the observed fraction of violations in the sequence

Unconditional Coverage Testing

- Plugging the maximum likelihood (ML) estimates back into the likelihood function gives the optimized likelihood as

$$L\left(\hat{\pi}\right) = \left(1 - T_1/T\right)^{T_0} \left(T_1/T\right)^{T_1}$$

- Under the unconditional coverage null hypothesis that π=*p*, where *p* is the known *VaR* coverage rate, we have the likelihood

$$L\left(p\right) = \prod_{t=1}^{T} (1-p)^{1-I_{t+1}} p^{I_{t+1}} = \left(1-p\right)^{T_0} p^{T_1}$$

- We can check the unconditional coverage hypothesis using a likelihood ratio test

$$LR_{uc} = -2\ln\left[L\left(p\right)/L\left(\hat{\pi}\right)\right]$$

Unconditional Coverage Testing

- Asymptotically, that is, as the number of observations, *T,* goes to infinity, the test will be distributed as a χ^2 with one degree of freedom
- Substituting in the likelihood functions, we write

$$LR_{uc} = -2\ln\left[(1-p)^{T_0}\,p^{T_1}/\left\{(1-T_1/T)^{T_0}\,(T_1/T)^{T_1}\right\}\right] \sim \chi_1^2$$

- The larger the LR_{uc} value is the more unlikely the null hypothesis is to be true
- Choosing a significance level of say 10% for the test, we will have a critical value of 2.7055 from the $\chi^2{}_1$ distribution

Unconditional Coverage Testing

- If the LR_{uc} test value is larger than 2.7055, then we reject the *VaR* model at the 10% level

- Alternatively, we can calculate the P-value associated with our test statistic

- The P-value is defined as the probability of getting a sample that conforms even less to the null hypothesis than the sample we actually got given that the null hypothesis is true

Unconditional Coverage Testing

- In this case, the P-value is calculated as

$$\text{P-value} \equiv 1 - F_{\chi_1^2}\left(LR_{uc}\right)$$

- Where $F_{\chi_1^2}(*)$ denotes the cumulative density function of a χ^2 variable with one degree of freedom
- If the P-value is below the desired significance level, then we reject the null hypothesis
- If we, for example, obtain a test value of 3.5, then the associated P-value is

$$\text{P-value} = 1 - F_{\chi_1^2}(3.5) = 1 - 0.9386 = 0.0614$$

Unconditional Coverage Testing

- If we have a significance level of 10%, then we would reject the null hypothesis, but if our significance level is only 5%, then we would not reject the null that the risk model is correct on average
- The choice of significance level comes down to an assessment of the costs of making two types of mistakes:
- We could reject a correct model (Type I error) or we could fail to reject (that is, accept) an incorrect model (Type II error).
- Increasing the significance level implies larger Type I errors but smaller Type II errors and vice versa

Unconditional Coverage Testing

- In academic work, a significance level of 1%, 5%, or 10% is typically used
- In risk management, Type II errors may be very costly so that a significance level of 10% may be appropriate
- Often, we do not have a large number of observations available for backtesting, and we certainly will typically not have a large number of violations, T_1, which are the informative observations
- It is therefore often better to rely on Monte Carlo simulated P-values rather than those from the χ^2 distribution

Unconditional Coverage Testing

- The simulated P-values for a particular test value can be calculated by first generating 999 samples of random i.i.d. Bernoulli(*p*) variables, where the sample size equals the actual sample at hand.
- Given these artificial samples we can calculate 999 simulated test statistics, call them $\left\{\widetilde{LR}_{uc}(i)\right\}_{i=1}^{999}$
- The simulated P-value is then calculated as the share of simulated LR_{uc} values that are larger than the actually obtained LR_{uc} test value

Unconditional Coverage Testing

- We can write

$$\text{P-value} = \frac{1}{1000}\left\{1 + \sum_{i=1}^{999} \mathbf{1}\left(\widetilde{LR}_{uc}(i) > LR_{uc}\right)\right\}$$

- where **1(•)** takes on the value of one if the argument is true and zero otherwise
- To calculate the tests in the first place, we need samples where *VaR* violations actually occurred; that is, we need some ones in the hit sequence
- If we, for example, discard simulated samples with zero or one violations before proceeding with the test calculation, then we are in effect conditioning the test on having observed at least two violations

Independence Test

- We should be concerned if all of the VaR violations or "hits" in a sample are happening around the same time which was the case in Figure 13.1
- If the VaR violations are clustered then the risk manager can essentially predict that if today is a violations, then tomorrow is more than *p*.100% likely to be a violation as well. This is clearly not satisfactory.
- In such a situation the risk manager should increase the VaR in order to lower the conditional probability of a violation to the promised *p*
- Our task is to establish a test which will be able to reject VaR with clustered violations

Independence Test

- To this end, assume the hit sequence is dependent over time and that it can be described as a so-called first-order Markov sequence with transition probability matrix

$$\Pi_1 = \begin{bmatrix} 1-\pi_{01} & \pi_{01} \\ 1-\pi_{11} & \pi_{11} \end{bmatrix}$$

- These transition probabilities simply mean that conditional on today being a non-violation (that is $I_t=0$), then the probability of tomorrow being a violation (that is, $I_{t+1} = 1$) is π_{01}

Independence Test

- The probability of tomorrow being a violation given today is also a violation is defined by

$$\pi_{11} = \Pr\left(I_{t+1} = 1 | I_t = 1\right)$$

- Similarly, the probability of tomorrow being a violation given today is not a violation is defined by

$$\pi_{01} = \Pr\left(I_{t+1} = 1 | I_t = 0\right)$$

- The first-order Markov property refers to the assumption that only today’s outcome matters for tomorrow’s outcome
- As only two outcomes are possible (zero and one), the two probabilities π_{01} and π_{11} describe the entire process

Independence Test

- The probability of a nonviolation following a nonviolation is 1-π_{01}, and the probability of a nonviolation following a violation is 1-π_{11}

- If we observe a sample of T observations, the likelihood function of the first-order Markov process as

$$L(\Pi_1) = (1 - \pi_{01})^{T_{00}} \pi_{01}^{T_{01}} (1 - \pi_{11})^{T_{10}} \pi_{11}^{T_{11}}$$

- where T_{ij}, $i, j = 0,1$ is the number of observations with a j following an i

Independence Test

- Taking first derivatives with respect to π_{01} and π_{11} and setting these derivatives to zero, we can solve for the maximum likelihood estimates

$$\begin{aligned}\hat{\pi}_{01} &= \frac{T_{01}}{T_{00}+T_{01}} \\ \hat{\pi}_{11} &= \frac{T_{11}}{T_{10}+T_{11}}\end{aligned}$$

- Using then the fact that the probabilities have to sum to one, we have

$$\begin{aligned}\hat{\pi}_{00} &= 1-\hat{\pi}_{01} \\ \hat{\pi}_{10} &= 1-\hat{\pi}_{11}\end{aligned}$$

Independence Test

- which gives the matrix of estimated transition probabilities

$$\hat{\Pi}_1 \equiv \begin{bmatrix} \hat{\pi}_{00} & \hat{\pi}_{01} \\ \hat{\pi}_{10} & \hat{\pi}_{11} \end{bmatrix} = \begin{bmatrix} 1 - \hat{\pi}_{01} & \hat{\pi}_{01} \\ 1 - \hat{\pi}_{11} & \hat{\pi}_{11} \end{bmatrix} = \begin{bmatrix} \frac{T_{00}}{T_{00}+T_{01}} & \frac{T_{01}}{T_{00}+T_{01}} \\ \frac{T_{10}}{T_{10}+T_{11}} & \frac{T_{11}}{T_{10}+T_{11}} \end{bmatrix}$$

- Allowing for dependence in the hit sequence corresponds to allowing π_{01} to be different from π_{11}
- We are typically worried about positive dependence, which amounts to the probability of a violation following a violation (π_{11})being larger than the probability of a violation following a nonviolation (π_{01})

Independence Test

- If, on the other hand, the hits are independent over time, then the probability of a violation tomorrow does not depend on today being a violation or not, and we write $\pi_{01} = \pi_{11} = \pi$
- Under independence, the transition matrix is thus

$$\hat{\Pi} = \begin{bmatrix} 1-\hat{\pi} & \hat{\pi} \\ 1-\hat{\pi} & \hat{\pi} \end{bmatrix}$$

- We can test the independence hypothesis that $\pi_{01} = \pi_{11}$ using a likelihood ratio test

$$LR_{ind} = -2\ln\left[L\left(\hat{\Pi}\right)/L\left(\hat{\Pi}_1\right)\right] \sim \chi_1^2$$

Independence Test

- where $L\left(\hat{\Pi}\right)$ is the likelihood under the alternative hypothesis from the LR_{uc} test
- In large samples, the distribution of the LR_{ind} test statistic is also χ^2 with one degree of freedom
- But we can calculate the P-value using simulation as we did before
- We again generate 999 artificial samples of i.i.d. Bernoulli variables, calculate 999 artificial test statistics, and find the share of simulated test values that are larger than the actual test value.

Independence Test

- As a practical matter, when implementing the LR_{ind} tests we may incur samples where $T_{11} = 0$
- In this case, we simply calculate the likelihood function as

$$L\left(\hat{\Pi}_1\right) = \left(1 - \hat{\pi}_{01}\right)^{T_{00}} \hat{\pi}_{01}^{T_{01}}$$

Conditional Coverage Testing

- Ultimately, we care about simultaneously testing if the *VaR* violations are independent and the average number of violations is correct

- We can test jointly for independence and correct coverage using the conditional coverage test

$$LR_{cc} = -2\ln\left[L\left(p\right)/L\left(\hat{\Pi}_1\right)\right] \sim \chi_2^2$$

- which corresponds to testing that $\pi_{01} = \pi_{11} = p$

Conditional Coverage Testing

- Notice that the LR_{cc} test takes the likelihood from the null hypothesis in the LR_{uc} test and combines it with the likelihood from the alternative hypothesis in the LR_{ind} test.
- Therefore,

$$\begin{aligned}
LR_{cc} &= -2\ln\left[L(p)/L\left(\hat{\Pi}_1\right)\right] \\
&= -2\ln\left[\left\{L(p)/L\left(\hat{\Pi}\right)\right\}\left\{L(\hat{\pi})/L\left(\hat{\Pi}_1\right)\right\}\right] \\
&= -2\ln\left[L(p)/L\left(\hat{\Pi}\right)\right] - 2\ln\left[L\left(\hat{\Pi}\right)/L\left(\hat{\Pi}_1\right)\right] \\
&= LR_{uc} + LR_{ind}
\end{aligned}$$

Conditional Coverage Testing

- so that the joint test of conditional coverage can be calculated by simply summing the two individual tests for unconditional coverage and independence

- As before, the P-value can be calculated from simulation