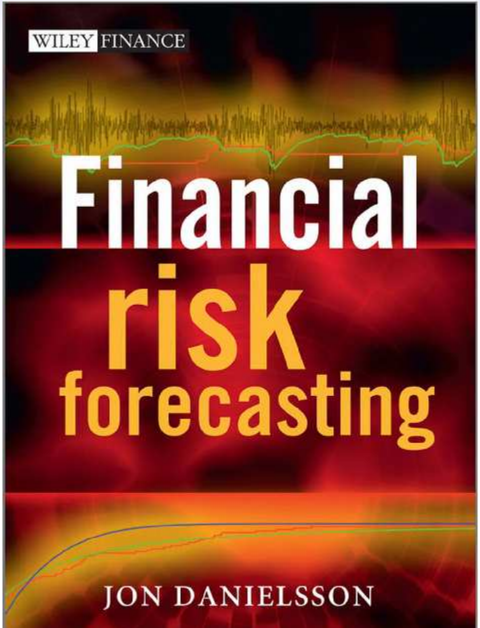
Financial Risk Forecasting
Chapter 8
Backtesting And Stresstesting

Jon Danielsson ©2023
London School of Economics

To accompany
Financial Risk Forecasting
www.financialriskforecasting.com
Published by Wiley 2011

Version 8.0, August 2023


WILEY FINANCE
Financial
risk
forecasting
JON DANIELSSON

Backtesting And Stresstesting

Introduction

- When making a risk forecast (or any type of forecast)
- It is important to validate the forecast

  Ex post: ideally this is done after we make them – using *operational criteria*

  Ex ante: but often we have to do it before
- VaRs are only observed infrequently, a long period of time would be required
- *Backtesting* evaluates VaR forecasts by checking how a VaR forecast model performs over a period in the past – *in-sample*

It's Not Perfect

- While the idea of a backtest sounds good in theory, there are serious issues in practice
- We will return to this later after we have covered backtesting
- But the basic problem is that the person doing the backtest knows the future
- And therefore can *adjust* the forecast to perform *too well*

The Focus of This Chapter

- Backtesting
- Application of backtesting
- Significance of backtests
  - Bernoulli coverage test
  - Testing the independence of violations
  - Joint test
  - Loss-function-based backtests
- Expected shortfall backtesting
- Problems with backtesting
- Stress testing

Notation

W_T	Testing window size
W_E	Estimation window size
$T = W_E + W_T$	Number of observations in a sample
η	Indicates whether a violation occurs
υ	Count of violations

Backtesting

What Is Backtesting?

- Procedure to compare various risk models, *ex ante* (that is in-sample)
- Take ex ante VaR forecasts from a particular model and compare them with *ex post* realised return (that is, historical observations)
- Whenever losses exceed VaR, a *VaR violation* is said to have occurred
- Can analyse violations in various ways

Machine Learning Comparison

- Learn to forecast risk out-of-sample in a training sample
- Evaluate model in testing sample
- Conceptually similar to what we do here
- Except we use specific models instead of (mostly) unsupervised learning
- When we know a lot about underlying stochastic process, will perform better
- Especially when samples are as small as in our case

Forecasting VaR: Example

- Imagine you have ten years of data, from 2014 to 2023
- And using the first two years of that
- To forecast risk for 1 January 2016

Forecasting VaR: Example (Cont.)

- The 500 trading days in 2014 and 2015 constitute the first *estimation window*
- W_E is then moved up by one day to obtain the risk forecast for the second day of 2014, etc.

Start	End	VaR forecast
1/1/2014	31/12/2015	VaR(1/1/2016)
2/1/2014	1/1/2016	VaR(2/1/2016)
⋮	⋮	⋮
31/12/2022	30/12/2023	VaR(31/12/2023)

Usefulness of Backtesting

- Identifying the weaknesses of risk forecasting methods
- Hence providing avenues for improvement
  - Not very informative about the *causes* of weaknesses
- Models that perform poorly during backtesting should question
  1. Model assumptions
  2. Parameter estimates
- Backtesting can prevent underestimation and overestimation of risk

Definitions

Estimation window (W_E): the number of observations used to forecast risk; if different procedures or assumptions are compared, the estimation window is set to whichever one needs the highest number of observations

Testing window (W_T): the data sample over which risk is forecast (that is, the days where we have made a VaR forecast)

$$T = W_E + W_T$$

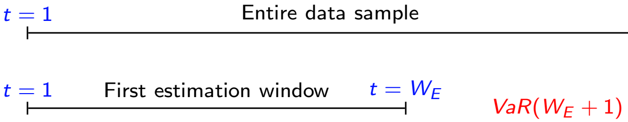
Dates and Indices

- VaR forecasts can be compared with the actual outcome
- The daily 2014 to 2023 returns are *already known*
- Instead of referring to calendar dates (for example, 1/1/2014), refer to days by indexing the returns, assuming *250 trading days* per year:
  - y_1 is the return on 1/1/2014
  - $y_{2,500}$ is the return on the last day, 31/12/2023

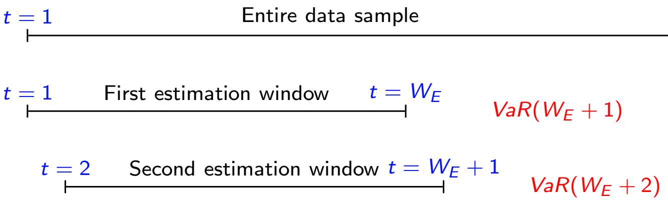
Testing

$t = 1$ Entire data sample $t = T$

Testing


$t = 1$
Entire data sample
$t = T$
$t = 1$
First estimation window
$t = W_E$
$VaR(W_E + 1)$

Testing


$t = 1$
Entire data sample
$t = T$
$t = 1$
First estimation window
$t = W_E$
$VaR(W_E + 1)$
$t = 2$
Second estimation window
$t = W_E + 1$
$VaR(W_E + 2)$

Testing

$t = 1$ Entire data sample $t = T$

$t = 1$ First estimation window $t = W_E$ $VaR(W_E + 1)$

$t = 2$ Second estimation window $t = W_E + 1$ $VaR(W_E + 2)$

$t = 3$ Third estimation window $t = W_E + 2$ $VaR(W_E + 3)$

Testing

$t = 1$ Entire data sample $t = T$

$t = 1$ First estimation window $t = W_E$ $VaR(W_E + 1)$

$t = 2$ Second estimation window $t = W_E + 1$ $VaR(W_E + 2)$

$t = 3$ Third estimation window $t = W_E + 2$ $VaR(W_E + 3)$

$\vdots$

Testing

$t = 1$ Entire data sample $t = T$

$t = 1$ First estimation window $t = W_E$ $VaR(W_E + 1)$

$t = 2$ Second estimation window $t = W_E + 1$ $VaR(W_E + 2)$

$t = 3$ Third estimation window $t = W_E + 2$ $VaR(W_E + 3)$

$\vdots$

$t = T - W_E$ Last estimation window $t = T - 1$ $VaR(T)$

- The estimation window W_E is set at 500 days, and the testing window W_T is therefore 2,000 days

t	$t + W_E - 1$	$VaR(t + W_E)$
1	*500*	*VaR(501)*
2	*501*	*VaR(502)*
$\vdots$	$\vdots$	$\vdots$
1,999	*2,499*	*VaR(2,500)*

Violation Ratios

VaR Violation

- If a financial loss on a particular day exceeds the VaR forecast, then the *VaR limit is said to have been violated*

VaR violation: an event such that

$$\eta_t = \begin{cases} 1, & \text{if } y_t \leq -\text{VaR}_t \\ 0, & \text{if } y_t > -\text{VaR}_t. \end{cases}$$

Counting Violations

- Count the violations

$$v_1$$

and non-violations

$$v_0$$

$$v_1 = \sum_{t=1}^{W_T} \eta_t$$

$$v_0 = W_T - v_1$$

Violation Ratios

- The main tools used in backtesting are *violation ratios*
- The *observed* number of VaR violations are compared with the *expected*

Violation ratio:

$$\text{VR} = \frac{\text{Observed number of violations}}{\text{Expected number of violations}} = \frac{v_1}{\rho \times W_T}$$

- If the violation ratio is greater than one, the VaR model *underforecasts* risk
- If smaller than one the model *overforecasts* risk

Estimation Window Length

- W_E determined by the choice of VaR model and probability level
- Different methods have different data requirements

EWMA	About 30 days
HS	At least 300 days for VaR(1%)
GARCH	500 or more days

Picking W_E

- The estimation window should be sufficiently large to accommodate the most stringent data criteria
- So if comparing EWMA and HS, use at least 300 for both
- Even within the same method, it may be helpful to compare different window lengths
- Maybe compare HS with 300, 500, and 1,000 days
- Or GARCH with 500 and 5,000 days

Testing Window Length

- VaR violations are infrequent events
- With a 1% VaR, a violation is expected once every 100 days, so that 2.5 violations are expected per year
- So the actual sample size of violations is quite small
- Causing difficulties for statistical inference
- At least 10 violations for reliable statistical analysis, or four years of data
- Preferably more

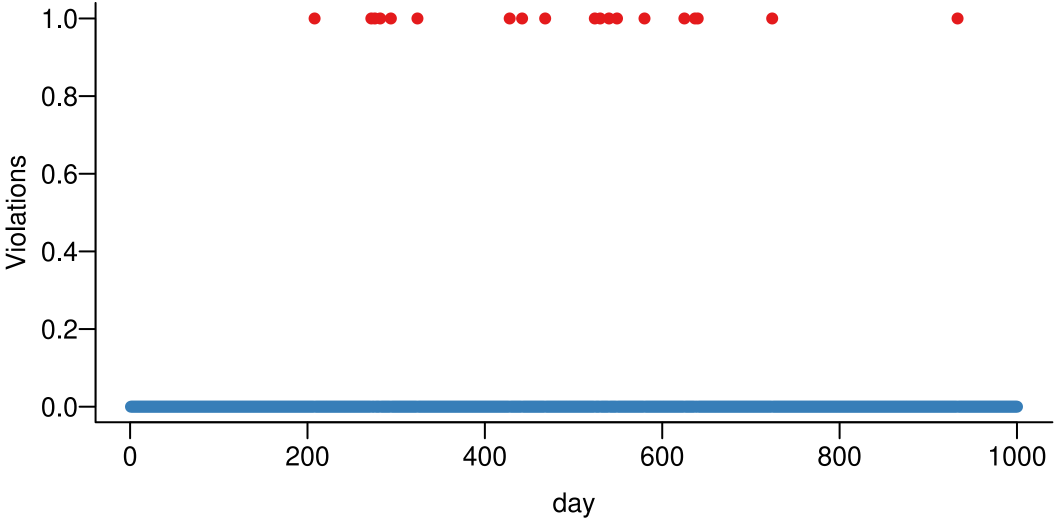
Violation Ratios

- *VR=1* is expected, but how can we ascertain whether any other value is statistically significant?
- A useful *rule of thumb*
  - If $VR \in [0.8, 1.2]$, the model is *good*
  - If $VR \in [0.5, 0.8]$, or $VR \in [1.2, 1.5]$, the model is *acceptable*
  - If $VR \in [0.3, 0.5]$, or $VR \in [1.5, 2]$, the model is *bad*
  - If $VR < 0.3$ or $VR > 2$ the model is *useless*
- Both bounds narrow with increasing testing window lengths
- As a first attempt
  - Plot the actual returns and VaR together
  - And then do a statistical test

Simulating Violations

- Suppose you want to simulate a coin toss in R
- `binom(prob=0.5,n=1,size=1)`
- Probability 50%, one observation and one try
- Suppose the VaR probability is 1% and we want to simulate a testing sample size of thousand days
- `rbinom(prob=0.01,n=1000,size=1)`

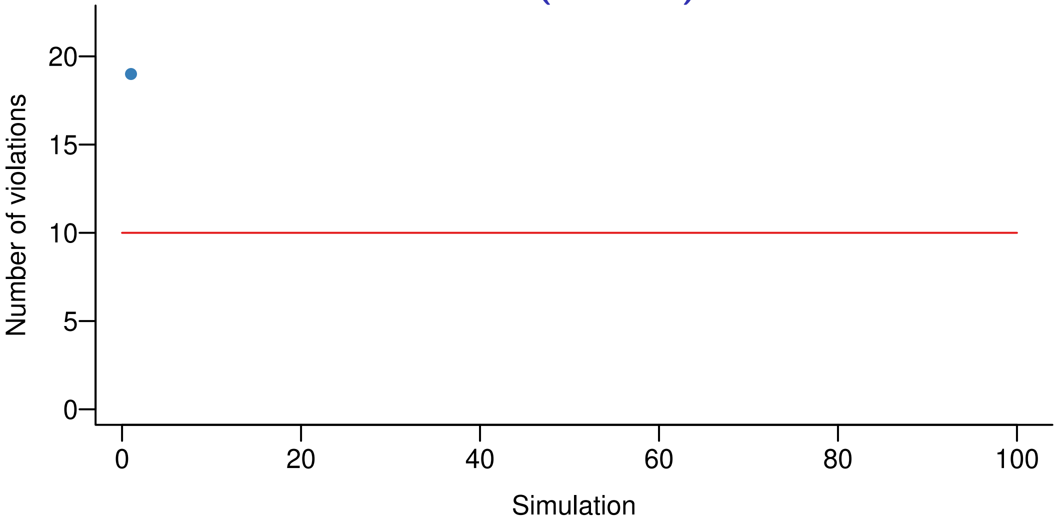
Outcome (Part I)



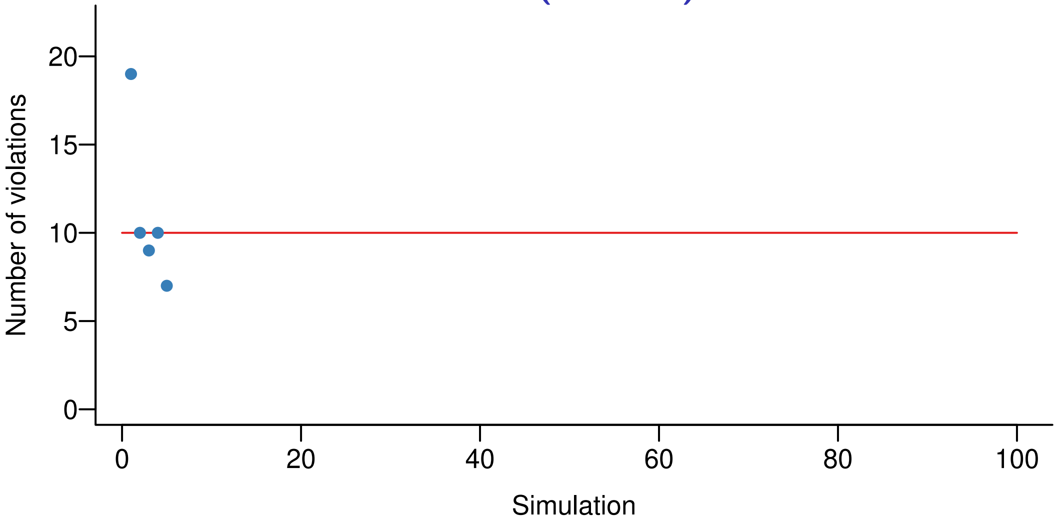
Outcome (Part II)

- And the number of violations
- `sum(rbinom(prob=0.01,n=1000,size=1))`
- Let's repeat that a few times

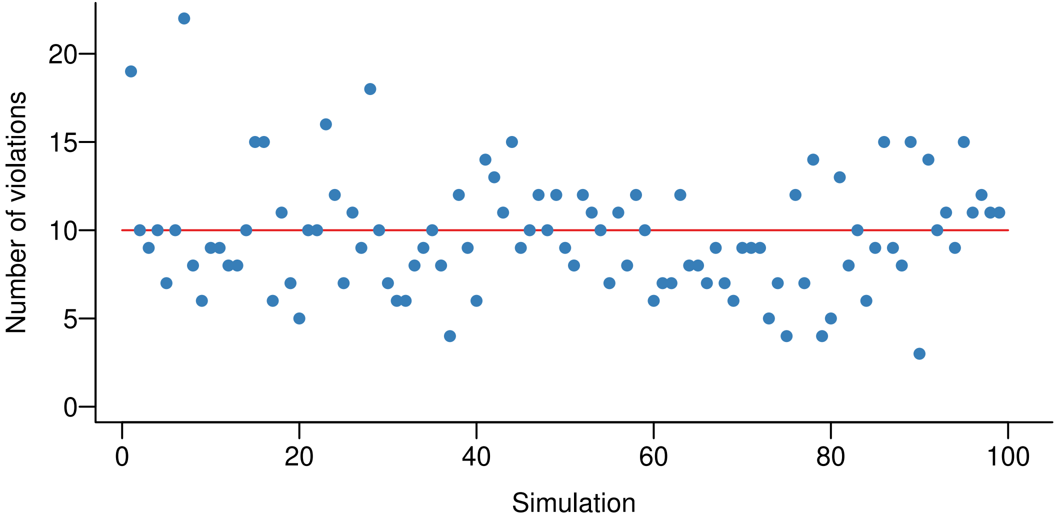
Outcome (Part III)



Outcome (Part IV)



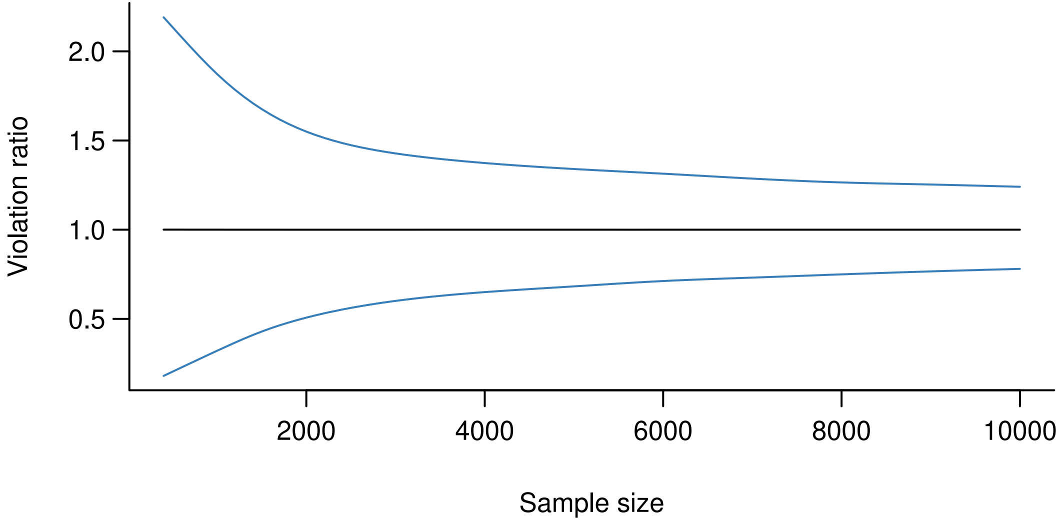
Outcome (Part V)



Simulation Estimation of Confidence Bounds

- By simulating a lot of times, we can construct Monte Carlo confidence bounds
- By taking the 0.5% and the 99.5% smallest violation ratios for each sample size
- We get the *empirical* 99% confidence bound

99% Empirical Confidence Bounds

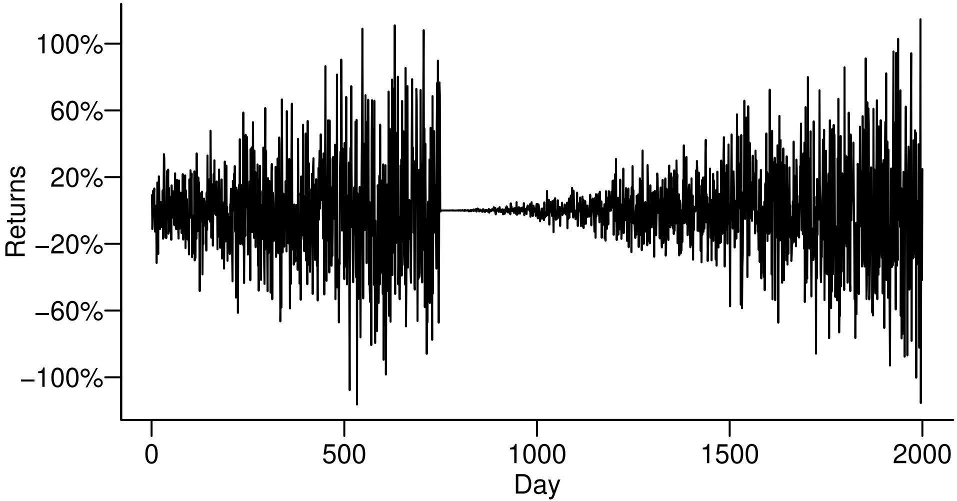


Application of Backtesting

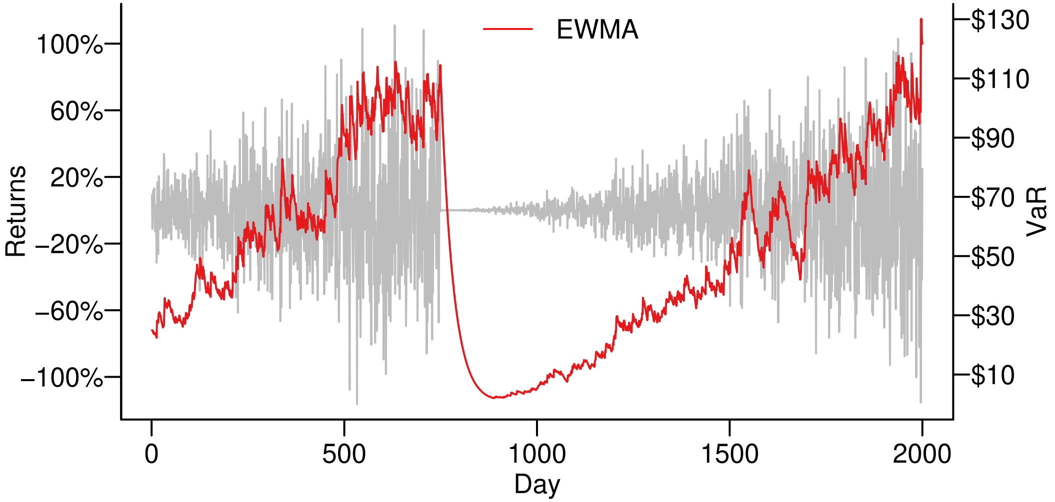
Extreme Example

- Start with extreme volatility clusters
- And pay a special attention to how the various methods react to the collapse of volatility to zero

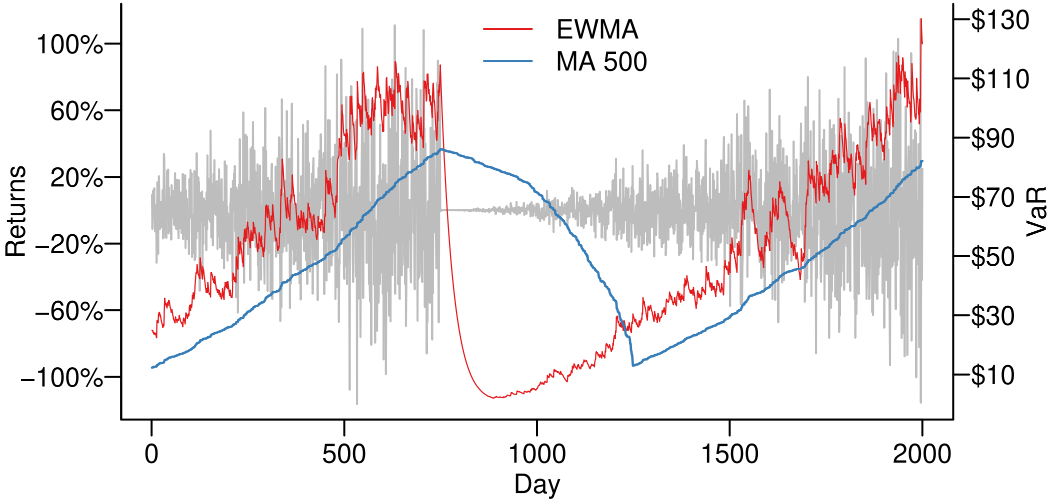
Volatility and VaR: Extreme Example



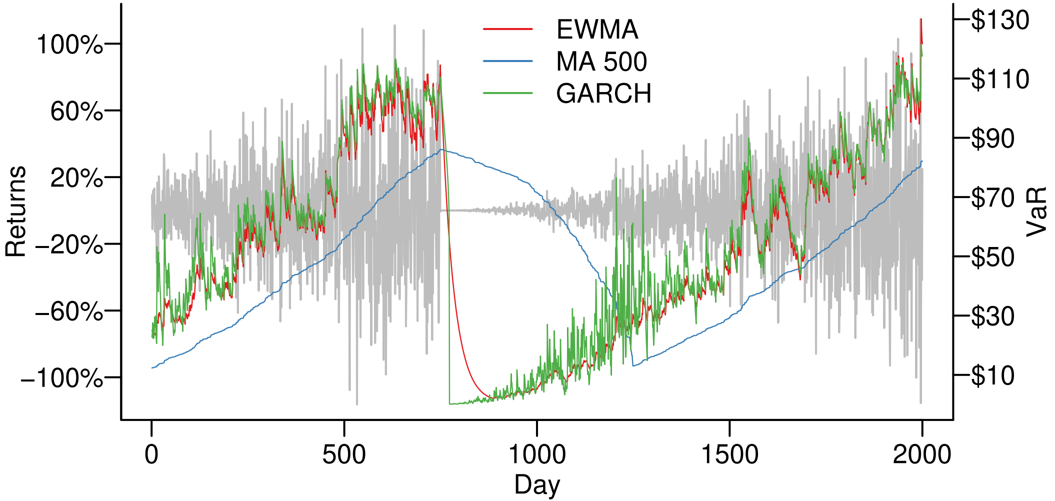
Volatility and VaR: Extreme Example



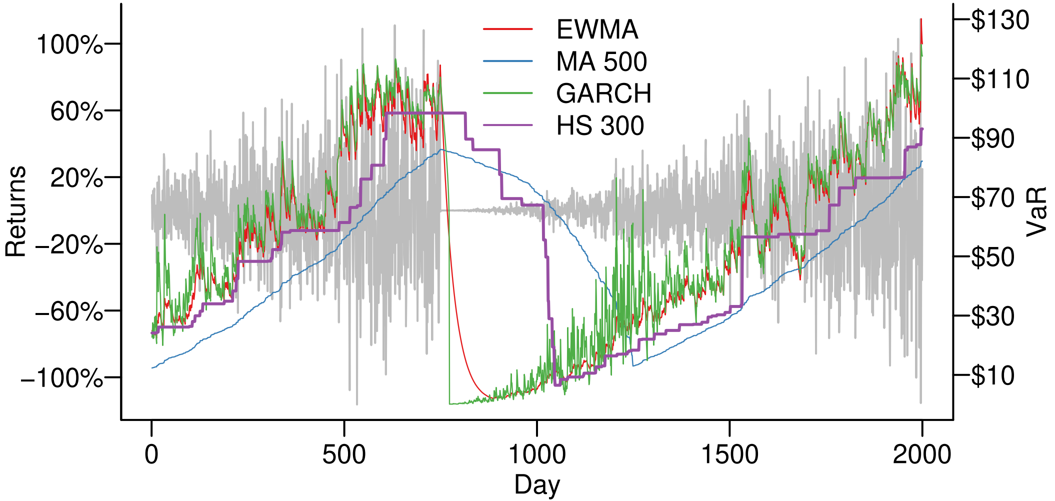
Volatility and VaR: Extreme Example



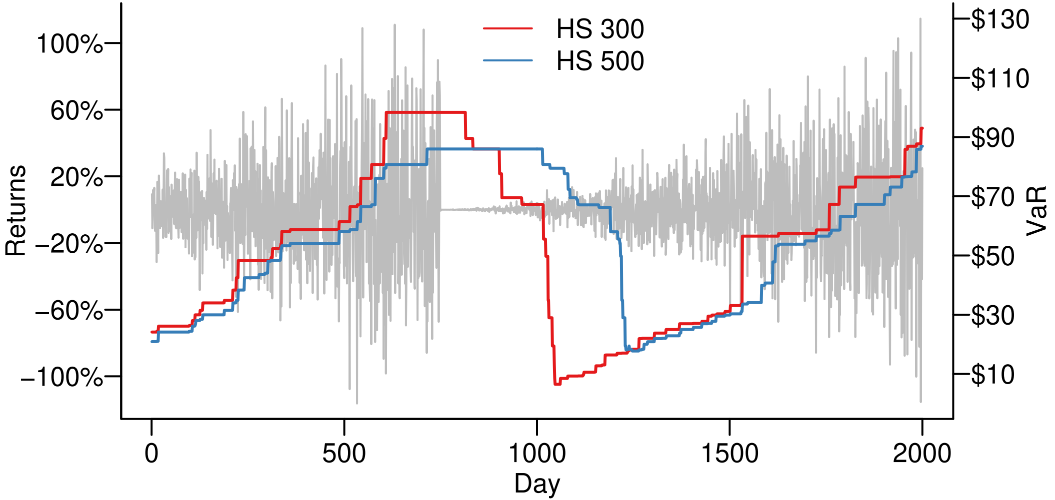
Volatility and VaR: Extreme Example



Volatility and VaR: Extreme Example



Volatility and VaR: Extreme Example



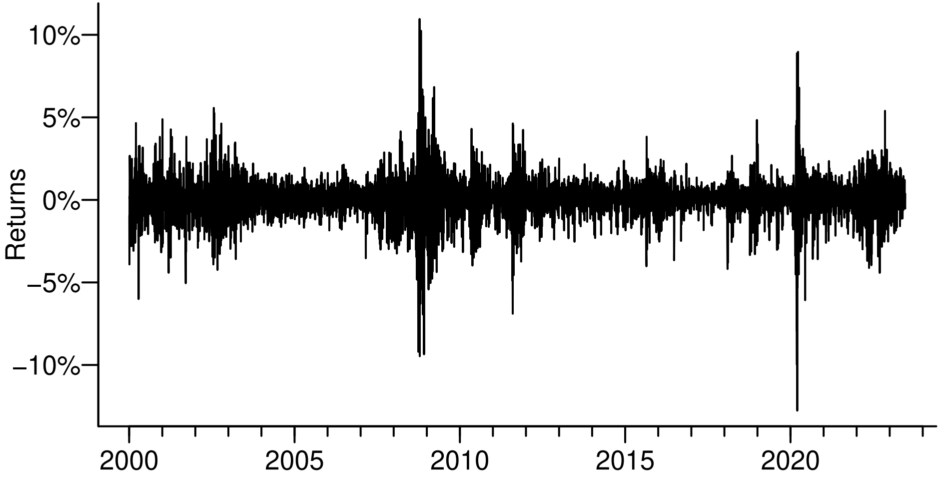
Summary Conclusion

- The worst performing model is MA as it is always behind
- Recall the discussion of the model in Chapter 2
- EWMA is usually quite close to GARCH, but GARCH is more noisy right after the volatility collapsed to zero
- The main reason is that half the sample is in high volatility environment and half in the low

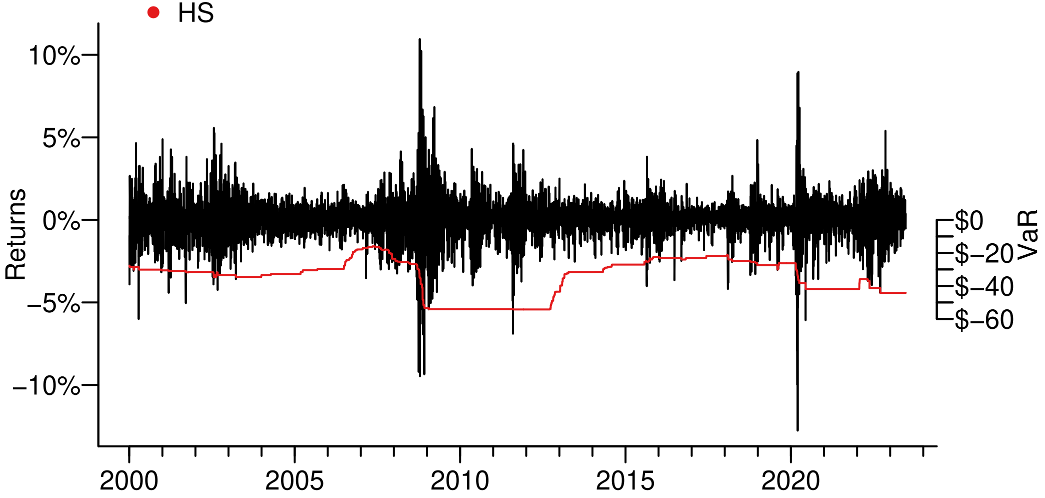
S&P-500

- We then take a sample from the S&P-500
- There are some especially interesting regimes,
- The collapse of volatility after 2003
- Like 2008 and 2020

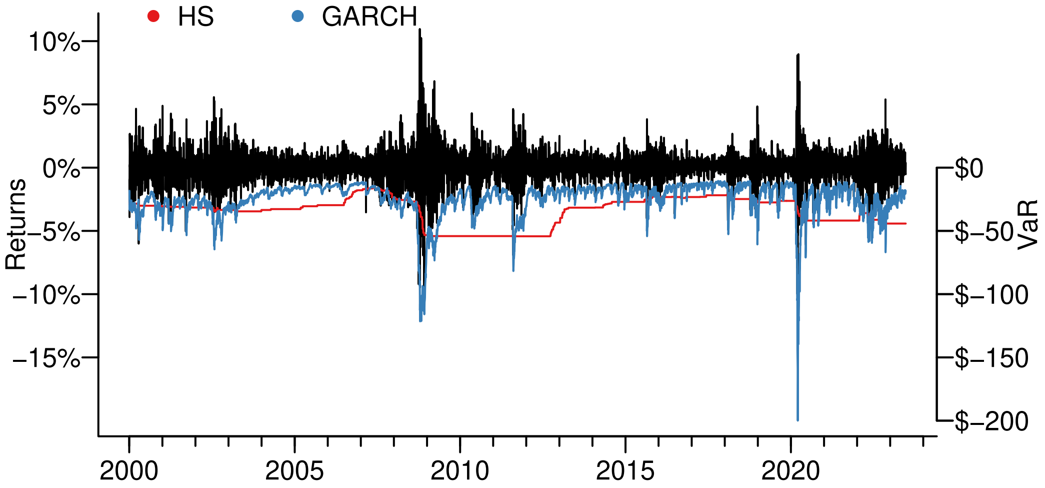
S&P-500 2000 to June 2023



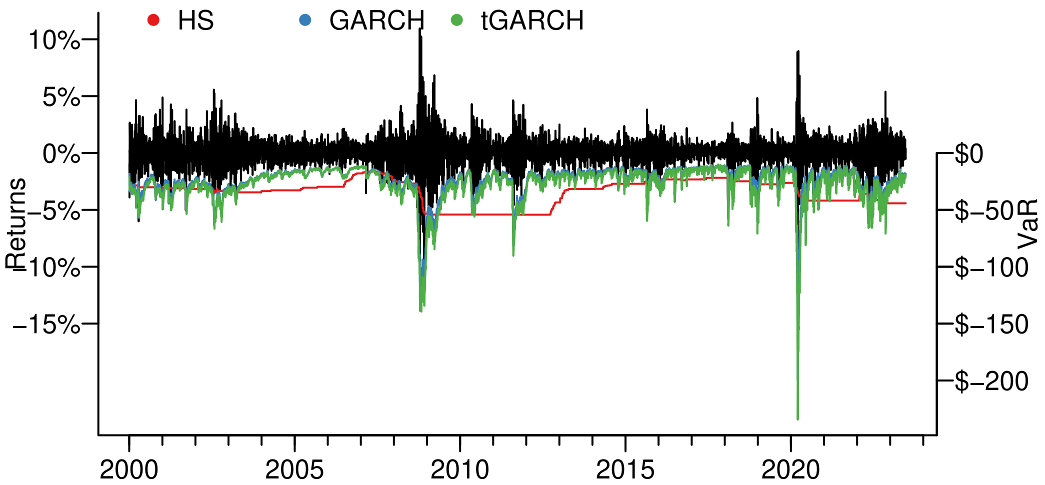
S&P-500 2000 to June 2023



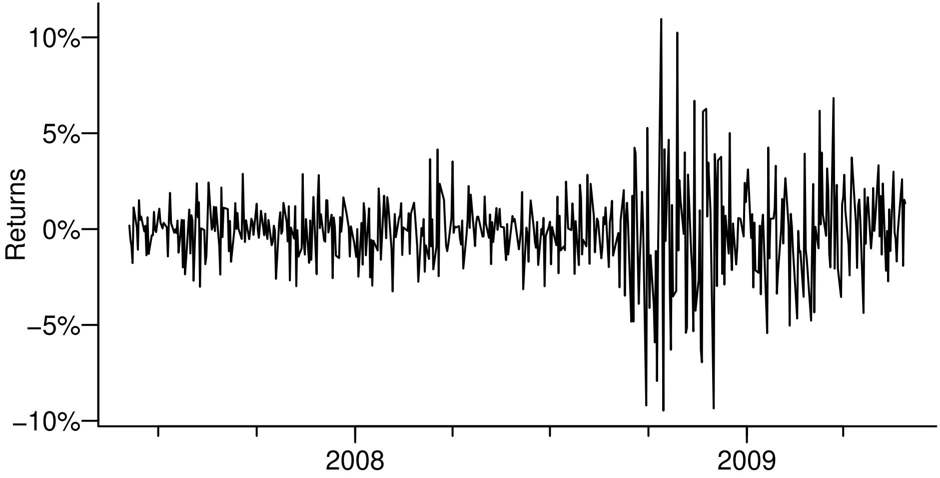
S&P-500 2000 to June 2023



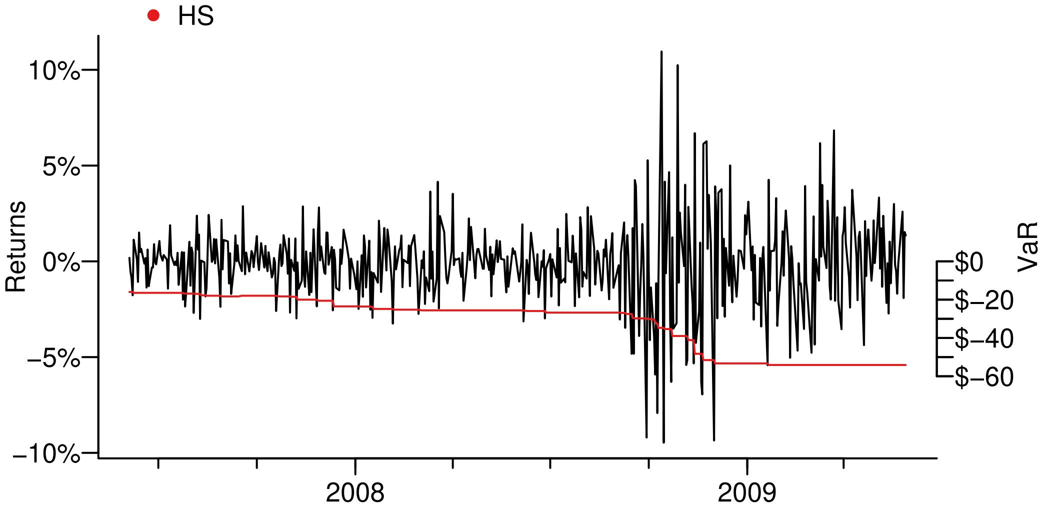
S&P-500 2000 to June 2023



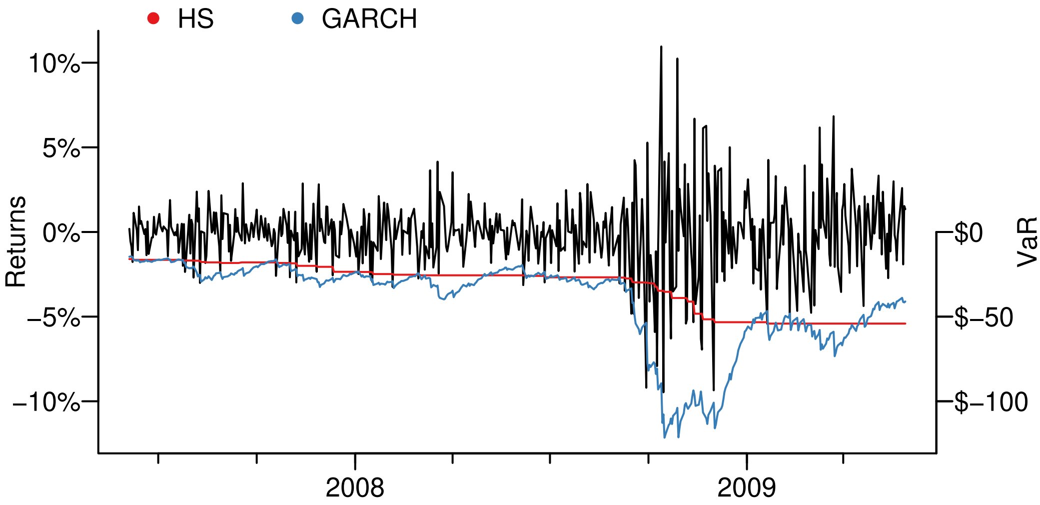
Zoom Into 2007 to 2009 Crisis



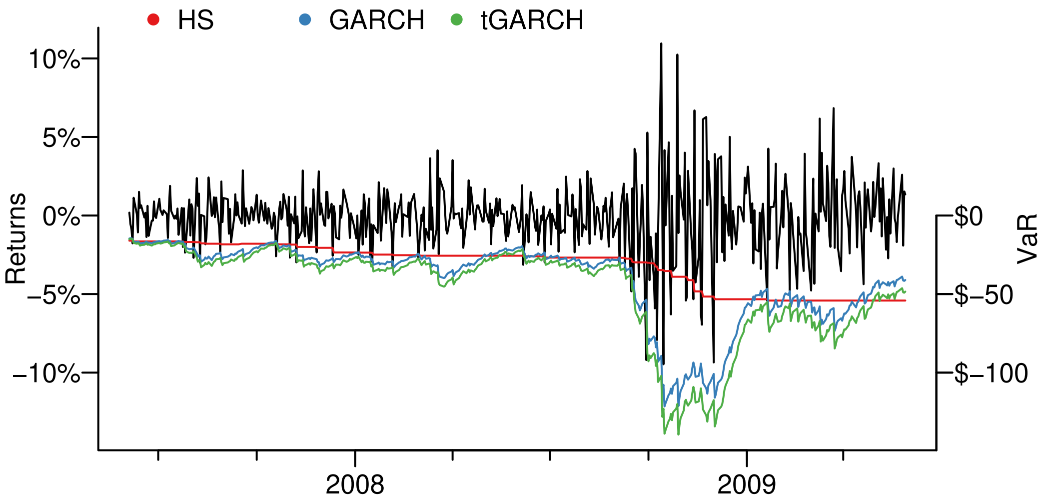
Zoom Into 2007 to 2009 Crisis



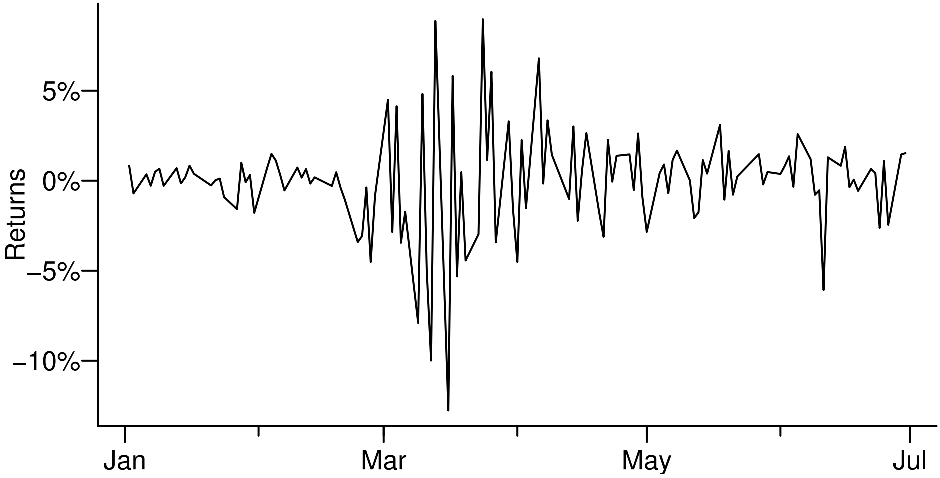
Zoom Into 2007 to 2009 Crisis



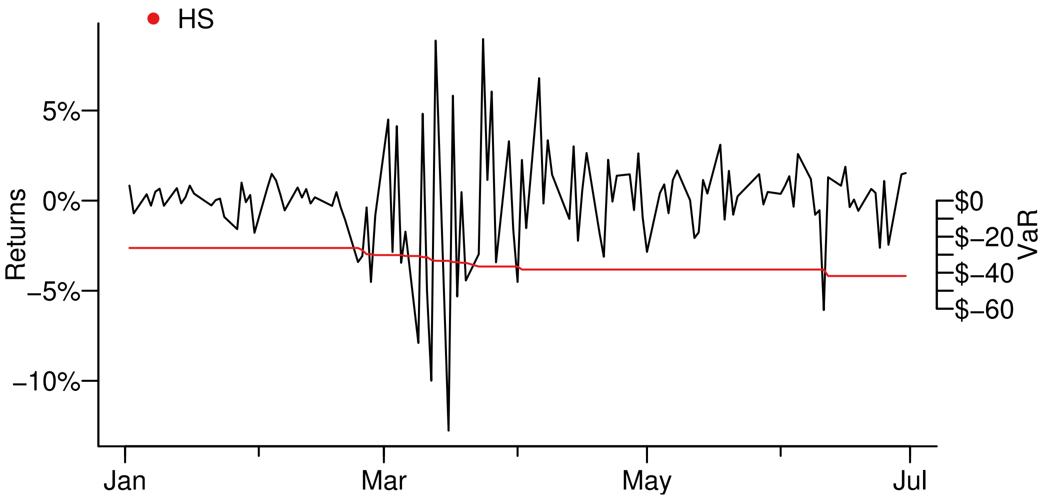
Zoom Into 2007 to 2009 Crisis



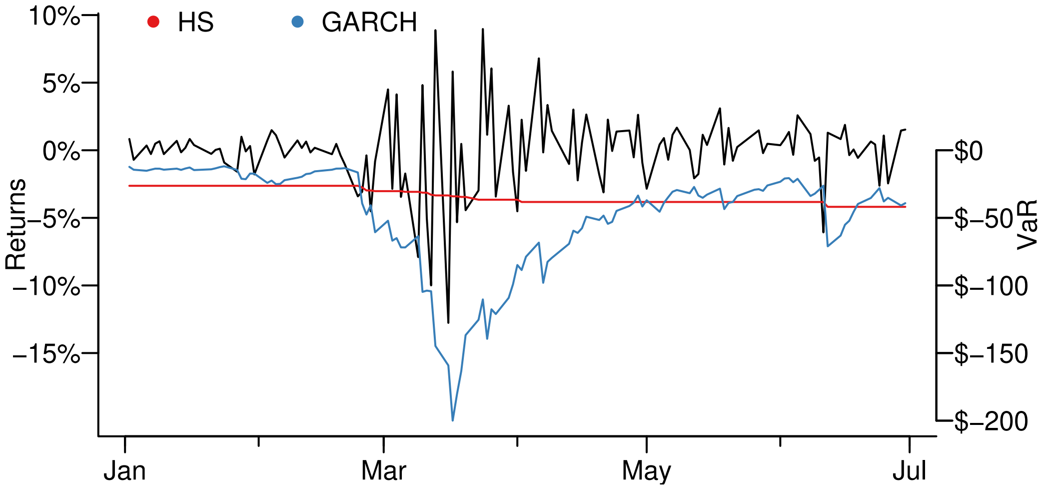
Zoom Into Covid



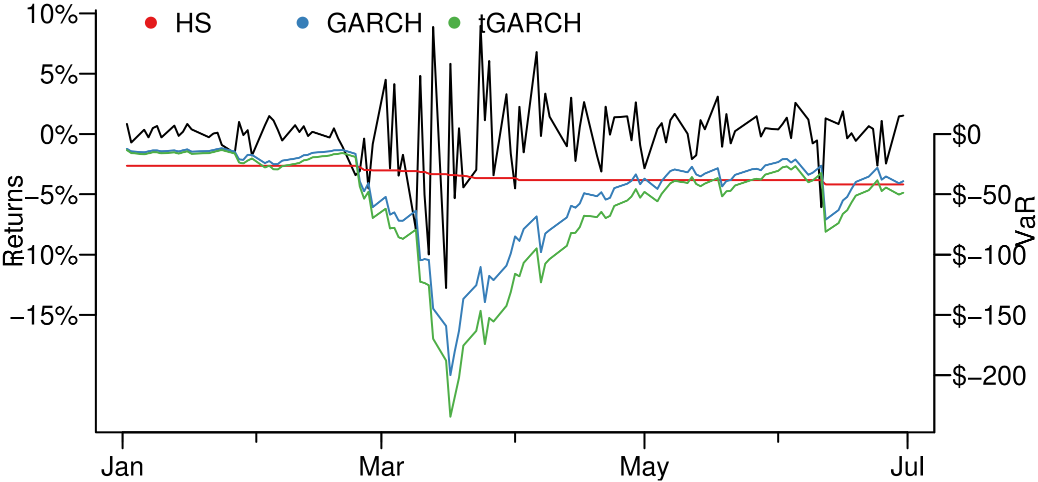
Zoom Into Covid



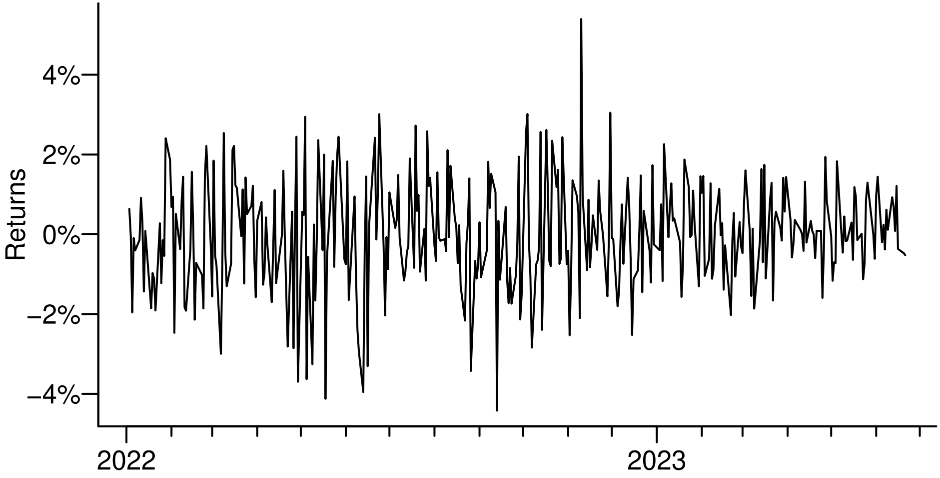
Zoom Into Covid



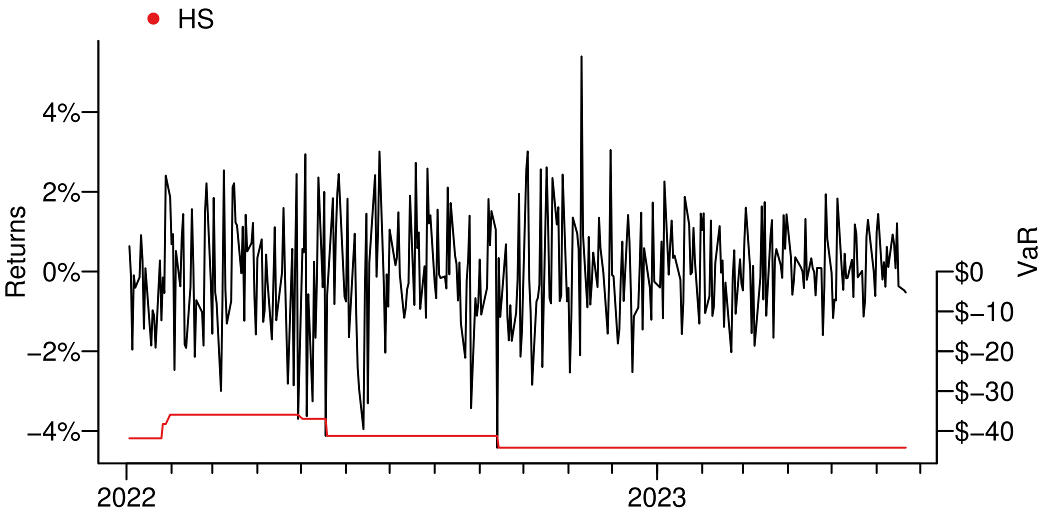
Zoom Into Covid



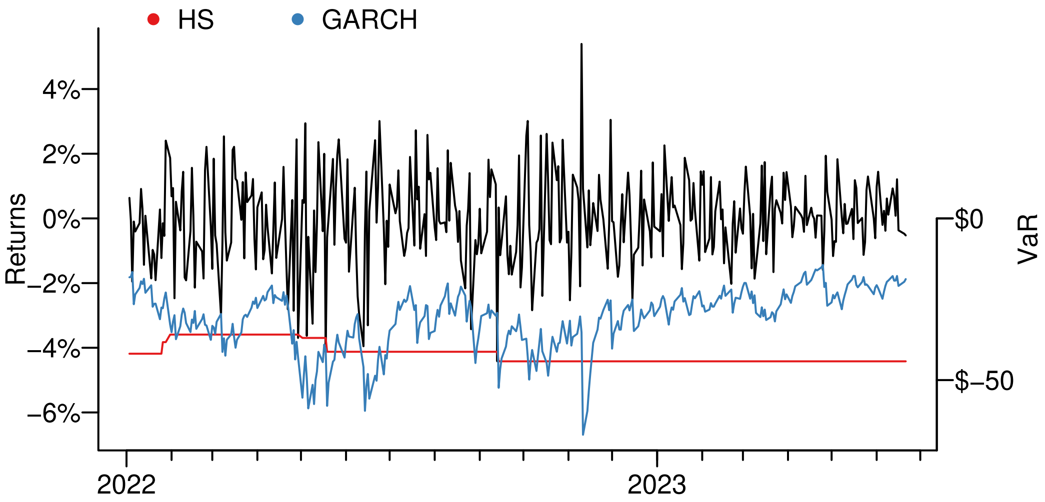
Zoom Into Russia - Ukraine War and Inflation



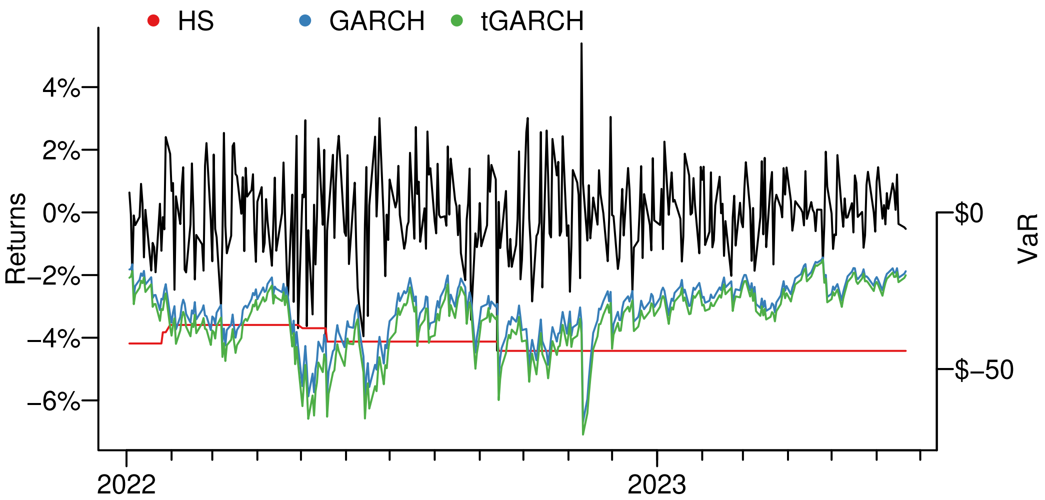
Zoom Into Russia - Ukraine War and Inflation



Zoom Into Russia - Ukraine War and Inflation



Zoom Into Russia - Ukraine War and Inflation



Significance of Backtests

Testing Violations

- We can test whether we get the expected number of violations and if there are patterns in the violations enumerate
  1. The number of violations (tested by the unconditional coverage)
  2. Clustering (tested by independence tests)

Distribution of Violations

- We have a sequence of returns, VaR and violations η_t

η	1	0	1	0	0	1	0	0	0	0
VaR	1.9	2.0	2.1	2.0	2.1	2.2	2.3	2.2	2.3	2.4
y	-2.1	1.4	-5.2	2.3	0.4	-3.7	4.1	0.1	3.2	-0.2
days	1	2	3	4	5	6	7	8	9	10

- The $\{\eta_t\}_{t=W_E+1}^{T}$ is a sequence of 1 or 0
- And hence follows the Bernoulli distribution
- Note that the sequence starts at $W_E + 1$ and ends at T and is hence W_T long

Estimation

- The sample probability, $\hat{\rho}$, can be estimated by the average number of violations

$$\hat{\rho} = \frac{v_1}{W_T}$$

- The Bernoulli density (on day t) is given by:

$$(1-\rho)^{1-\eta_t}(\rho)^{\eta_t}, \quad \eta_t = 0, 1$$

Bernoulli coverage test

Unconditional Coverage

- Does the expected number of violations, as given by ρ match the observed number of violations?
  - For a VaR(1%) backtest, we would expect to observe a violation 1% of the time
  - If, violations are observed more often, the VaR model is *underestimating risk*
  - And similarly if we observe too few violations

Bernoulli Coverage Test

- We can therefore test if the sequence $\{\eta_t\}_{t=W_E+1}^{T}$ has the expected number of 1 and 0
- Use the Bernoulli coverage test
- The null hypothesis for VaR violations is:

$$H_0 : \eta \sim B(\rho),$$

where B stands for the Bernoulli distribution

- Recall from Chapter 2 that the likelihood function is the product of the densities, and therefore
- The likelihood function is given by:

$$L_U(\hat{\rho}) = \prod_{t=W_E+1}^{T} (1-\hat{\rho})^{1-\eta_t}(\hat{\rho})^{\eta_t} = (1-\hat{\rho})^{\upsilon_0}(\hat{\rho})^{\upsilon_1}$$

- Denote this as the unrestricted likelihood function
- Because it uses estimated probability $\hat{\rho}$
- Under H_0, $\rho = \hat{\rho}$, so the restricted likelihood function is:

$$\begin{aligned}\mathcal{L}_R(\rho) &= \prod_{t=W_E+1}^{T} (1-\rho)^{1-\eta_t}(\rho)^{\eta_t} \\ &= (1-\rho)^{\upsilon_0}(\rho)^{\upsilon_1}\end{aligned}$$

- We can use a likelihood ratio (LR) test to see whether $\mathcal{L}_R = \mathcal{L}_U$ or, equivalently, whether $\rho = \hat{\rho}$:

$$
\begin{aligned}
\text{LR} &= 2(\log \mathcal{L}_U(\hat{\rho}) - \log \mathcal{L}_R(\rho)) \\
&= 2 \log \frac{(1-\hat{\rho})^{\upsilon_0}(\hat{\rho})^{\upsilon_1}}{(1-\rho)^{\upsilon_0}(\rho)^{\upsilon_1}} \\
&\overset{\text{asymptotic}}{\sim} \chi^2_{(1)}
\end{aligned}
$$

- Choosing a 5% significance level for the test, the null hypothesis is rejected if $LR > 3.84$

R

```
qchisq(p=1-0.05,df=1)
 3.841459
```

Bernoulli Coverage Test

R

```
bern_test=function(p,v){
  lv=length(v)
  sv=sum(v)
  al=log(p)*sv+log(1-0)*(lv-sv)
  bl=log(sv/lv)*sv +log(1-sv/lv)*(lv-sv)
  return(-2*(al-bl))
 }
```

Independence Property

Distribution of Violations

- Suppose the violations cluster

η	0	1	1	1	0	0	0	0	0	0
VaR	2.0	1.9	2.1	2.2	2.1	2.0	2.3	2.2	2.3	2.4
y	1.4	-2.1	-5.2	-3.7	0.4	2.3	4.1	0.1	3.2	-0.2
days	1	2	3	4	5	6	7	8	9	10

- Then we are violating the independence property

Independence Test

- Do two violations follow each other?
- They should not because
- If they do, we can predict a violation today if there was one yesterday
- A good VaR model would have increased the VaR forecast following a violation

Testing the Independence of Violations

- The probabilities of two consecutive violations is

$$\rho_{11}$$

- The probability of a violation if there was no violation on the previous day

$$\rho_{01}, \rho_{10}, \rho_{00}$$

- More generally, the probability that:

$$\rho_{ij} = \mathbb{P}\left(\eta_t = j | \eta_{t-1} = i\right).$$

- Where i and j are either 0 or 1

Testing the Independence of Violations (Cont.)

- The violation process can be represented as a Markov chain with two states
- So the first order transition probability matrix is defined as:

$$\Pi_1 = \begin{pmatrix} 1-\rho_{01} & \rho_{01} \\ 1-\rho_{11} & \rho_{11} \end{pmatrix}$$

- The likelihood function is:

$$L_1(\Pi_1) = (1-\rho_{01})^{v_{00}} \rho_{01}^{v_{01}} (1-\rho_{11})^{v_{10}} \rho_{11}^{v_{11}} \tag{8.5}$$

- Where v_{ij} is the number of observations where j follows i

Likelihood function

$$\hat{\Pi}_1 = \begin{pmatrix} \frac{v_{00}}{v_{00}+v_{01}} & \frac{v_{01}}{v_{00}+v_{01}} \\ \frac{v_{10}}{v_{10}+v_{11}} & \frac{v_{11}}{v_{10}+v_{11}} \end{pmatrix}$$

- Under the null hypothesis of no clustering, the probability of a violation tomorrow does not depend on today being a violation
- Then $\rho_{01} = \rho_{11} = p$ and the transition matrix is simply:

$$\Pi_2 = \begin{pmatrix} 1-p & p \\ 1-p & p \end{pmatrix}$$

- And the ML estimate is:

$$\hat{\rho} = \frac{v_{01} + v_{11}}{v_{00} + v_{10} + v_{01} + v_{11}}$$

- so

$$\hat{\Pi}_2 = \begin{pmatrix} 1-\hat{\rho} & \hat{\rho} \\ 1-\hat{\rho} & \hat{\rho} \end{pmatrix}$$

- The likelihood function then is

$$L_2(\Pi_2) = (1-p)^{v_{00}+v_{10}}\, p^{v_{01}+v_{11}} \tag{8.6}$$

Likelihood Ratio Test

- In (8.6) we impose independence but do not in (8.5)
- Replace the Π by the estimated numbers, $\hat{\Pi}$
- The LR test is then:

$$LR = 2\left(\log L_1\left(\hat{\Pi}_1\right) - \log L_2\left(\hat{\Pi}_2\right)\right) \overset{\text{asymptotic}}{\sim} \chi^2_{(1)}$$

Problems With the Independence Test

- The main problem with tests of this sort is that they must specify the particular way in which independence is breached
- However, there are many possible ways in which the independence property is not fulfilled:
  - Is the violation on days 1,3,5, and 7?
  - Test can't detect violation clustering

Testing the S&P-500

Testing S&P-500 1998 to 2009

Model	Coverage test		Independence test	
	Test statistic	p-value	Test statistic	p-value
EWMA	18.1	0.00	0.00	0.96
MA	81.2	0.00	7.19	0.01
HS	24.9	0.00	4.11	0.04
GARCH	16.9	0.00	0.00	0.99

1998 to 2006

Model	Coverage test		Independence test	
	Test statistic	p-value	Test statistic	p-value
EWMA	2.88	0.09	0.68	0.41
MA	6.15	0.01	2.62	0.11
HS	0.05	0.82	1.52	0.22
GARCH	1.17	0.28	0.99	0.32

Joint Test

- We can jointly test

$$\text{LR(joint)} = \text{LR(coverage)} + \text{LR(independence)} \sim \chi^2_{(2)}$$

- The joint test has less power to reject a VaR model which only satisfies one of the two properties

Expected Shortfall Backtesting

Expected Shortfall Backtesting

- It is harder to backtest expected shortfall (ES) than VaR because we are testing an *expectation* rather than a single *quantile*
- We know if VaR is violated, but cannot know that for ES
- There exists a simple methodology for backtesting ES that is analogous to the use of violation ratios for VaR
- For days when VaR is violated, normalised shortfall NS is calculated as:

$$NS_t = \frac{y_t}{\mathrm{ES}_t}$$

where ES_t is the observed ES on day t

Backtesting

- From the definition of ES, the expected Y_t given VaR is violated, is:

$$\frac{\mathrm{E}[Y_t | Y_t < -\mathrm{VaR}_t]}{\mathrm{ES}_t} = 1$$

- Therefore, average NS, $\overline{NS}$, should be one

$$H_0 : \overline{NS} = 1$$

Issues

- The reliability of any ES backtest procedure is much lower than that of VaR
  - With ES, we are testing whether the mean of returns on days when VaR is violated is the same as average ES on these days.
  - Much harder to create formal tests to ascertain whether normalised ES equals one or not than the coverage tests developed above for VaR violations
- Hence, backtesting ES requires many more observations than backtesting VaR
- In instances where ES is obtained directly from VaR, and gives the same signal as VaR (that is, when VaR is subadditive), it is better to simply use VaR

Problems with Backtesting

Structural Breaks

- Backtesting assumes that there have been *no structural breaks* in the data throughout the testing period:
    - But financial markets are continually evolving,
    - New technologies, assets, markets and institutions affect the statistical properties of market prices
    - Unlikely that the statistical properties of market data in the 1990s are the same as today,
    - Implying that a risk model that worked well then might not work well today

Intellectual Integrity

- Backtesting is only statistically valid if we have *no ex ante knowledge* of the data in the testing window
- If we iterate the process, continually refining the risk model with the same test data
  - and thus learning about the events in the testing window,
  - the model will be fitted to those particular outcomes,
  - violating underlying statistical assumptions
- So the actual confidence bounds are *wider* that suggested by the testing

Stresstesting

Stresstesting

- Create artificial market outcomes to see how risk management systems and risk models cope with the artificial event
- Assess the ability of a bank to survive a large shock
- The main aim is to come up with scenarios that are not well represented in recent historical data but are nonetheless possible and detrimental to portfolio performance

Examples of Historical Scenarios

Scenario	Period
Stock market crash	October 1987
Asian currency crisis	Summer 1997
LTCM and Russia crisis	August 1998
Global crisis	2007 to 2009
Eurozone crisis	Since 2010
Brexit	2017
Covid-19	2020

- Two types:
  1. Shocks that have never occurred or are more likely to occur than historical data suggest
  2. Shocks that reflect permanent or temporary structural breaks—where historical relationships do not hold

Stressed VaR

- Banks are now required to calculate stressed VaR
- While there are several ways to do that, here is a really simple approach
- Suppose we have a sample $1, \ldots, W_E, \ldots, T$
- We have a VaR_{t+1}
- The stressed VaR is

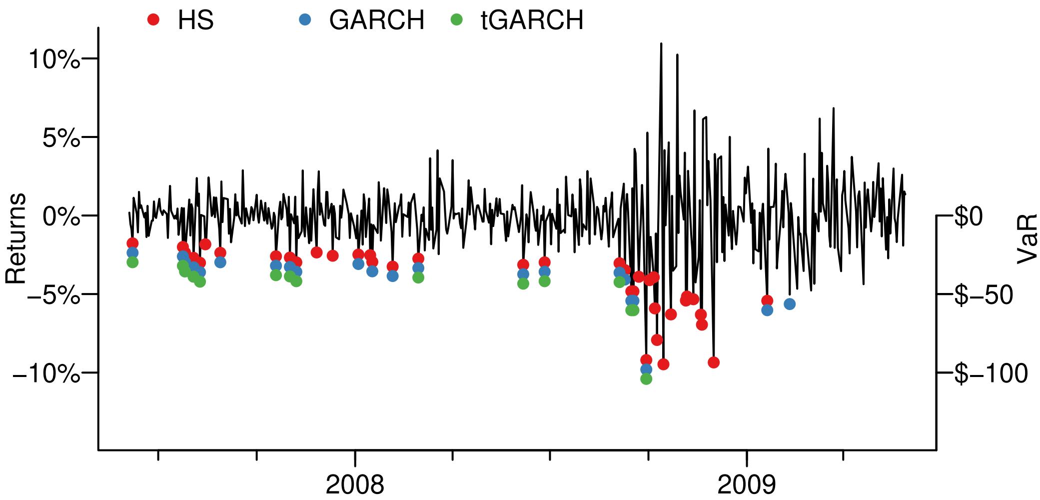
$$\text{SVaR}_{t+1} = \max \text{VaR}_i \; i = W_E + 1, \ldots, T + 1$$

Recent Stess Events

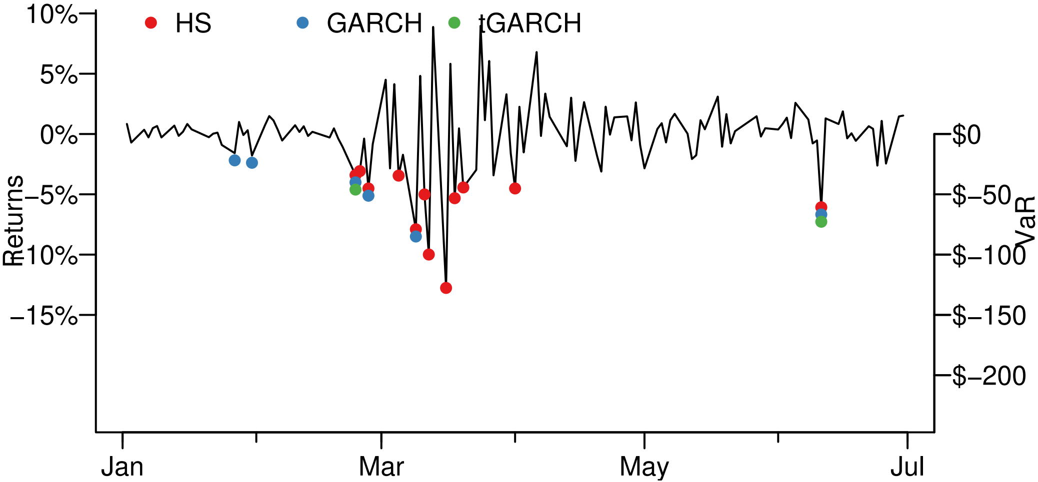
Backtesting the S&P-500 in Times of Stess

- Make the estimation window 1,000 days
- Testing window 1,000 days
- Probability: 1%
- Portfolio value one
- And compare GARCH and historical simulation
- We would expect 10 violations

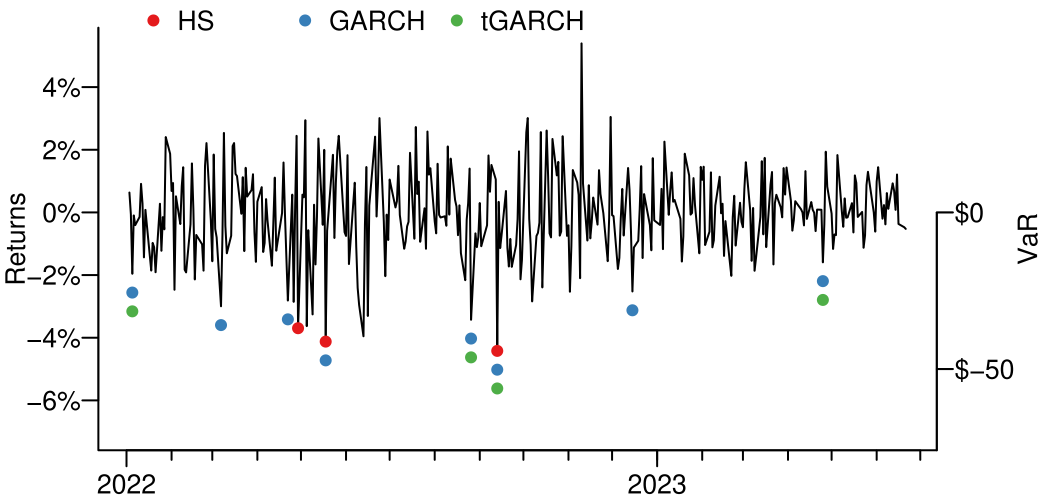
2007 to 2009 Crisis



Covid


HS
GARCH
tGARCH
10%
5%
0%
−5%
−10%
−15%
Returns
$0
$−50
$−100
$−150
$−200
VaR
Jan
Mar
May
Jul

Russia-Ukraine War to Inflation



Direct Comparison

- A direct comparison shows that most of the HS violations are at the height of the crisis
- While GARCH is more evenly distributed throughout the sample
- And interestingly is not violated on the worst day of the crisis
- Why do you think that is the case?
- So these results confirm what we have found for the same methods in other cases