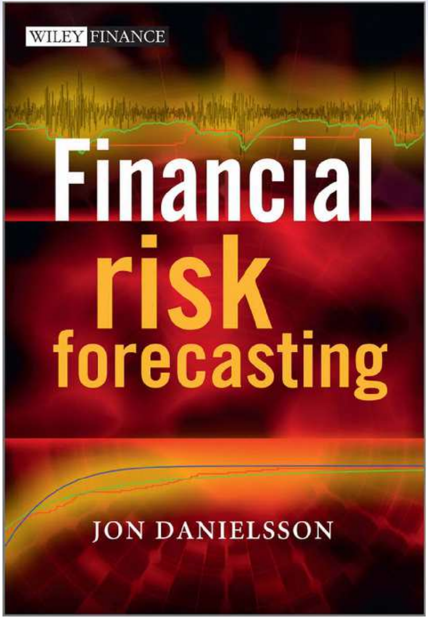
Financial Risk Forecasting
Chapter 8
Backtesting And Stresstesting

Jon Danielsson
London School of Economics

To accompany
Financial Risk Forecasting
`FinancialRiskForecasting.com`
Published by Wiley 2011

Version 10.0, August 2025


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:** validation after the forecast is made — using *operational criteria*
  - **Ex ante:** validation before observing outcomes — based on theory, design, or past performance
- 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*

Why Backtesting Isn't Foolproof

- 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
- The fundamental issue is that the analyst conducting the backtest has knowledge of future outcomes
- And therefore can *adjust* the forecast to perform *too well*
- In other words...
- The modeller knows the outcomes — and can tweak the forecast to perform suspiciously well

The Focus of This Chapter

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

Notations new to this chapter

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

Learning outcomes

1. Understand the difference between operational evaluation and backtesting
2. Know why backtesting is important
3. Know the concept of violations
4. Know the concept of the violation ratio
5. Understand how to implement tests of violation ratios
6. Be able to implement backtesting in R
7. Know the basic problems of backtesting ES

Backtesting

What Is Backtesting?

- Backtesting evaluates whether risk forecasts perform well out of sample
- We compare model-predicted VaR with actual realised returns
- 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)
- A loss exceeding VaR is called a *VaR violation* — it signals a forecast failure
- 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
- Unlike machine learning models, which are often data-driven and flexible, our risk models are structured and interpretable
- When we know a lot about underlying stochastic process, will perform better
- Especially when samples are as small as in our case
- If we know the underlying data-generating process, traditional models may outperform black-box methods

Elicitability: What Can Be Backtested?

- A risk measure is elicitable if it can be evaluated using a scoring (or loss) function
- This allows forecasters to be rewarded for accuracy and penalised for error
- VaR is elicitable:
    - A binary scoring rule: did the loss exceed the VaR threshold?
- ES is not elicitable on its own:
    - No scoring function uniquely incentivises correct ES forecasts
    - Makes standard backtesting and model comparison difficult

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)
$\vdots$	$\vdots$	$\vdots$
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 be questioned
  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

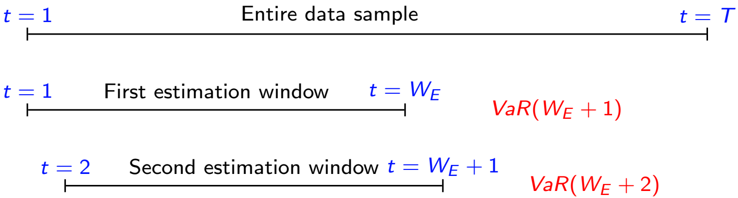
$t = 1$ Entire data sample $t = T$

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)
⋮	⋮	⋮
1,999	2,499	VaR(2,500)

What We have Set Up

- We now have a full set of VaR forecasts from a rolling model
- Each forecast corresponds to a specific historical day
- We will now compare forecasts to actual returns and evaluate accuracy
- Key concepts: violations, coverage, independence

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*
- Define the violation indicator η_t
- It equals 1 when the VaR is breached, 0 otherwise.

VaR violation: an event such that

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

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

- Over many forecasts, we expect a certain number of violations
- The *violation ratio* compares what we observed to what we expected
- 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

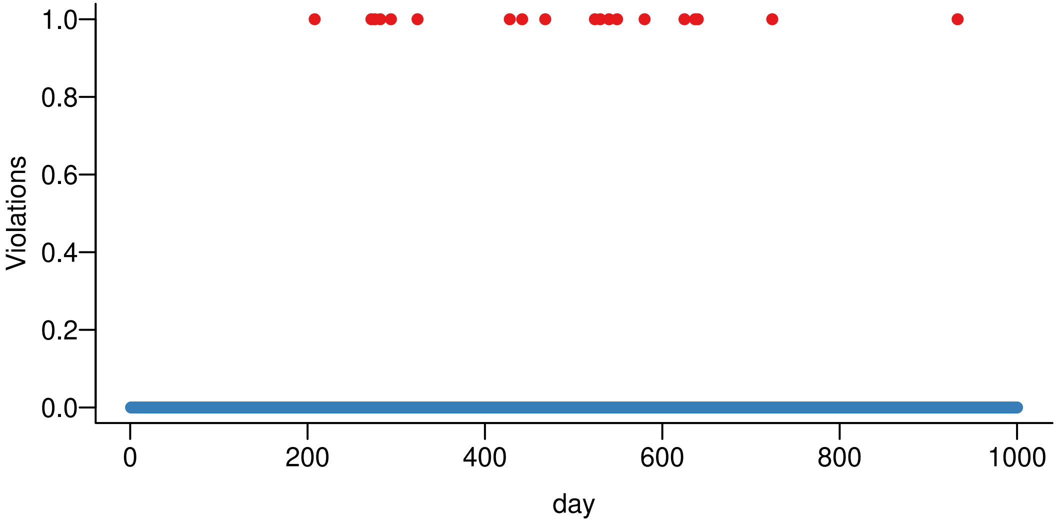
Simulating Violations

- Suppose you want to simulate a coin toss in R
- `rbinom(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)`

How Sample Size Affects the Violation Ratio

- Small samples $\rightarrow$ high variability in VR
- Larger samples $\rightarrow$ VR stabilises near 1
- Confidence bounds get narrower

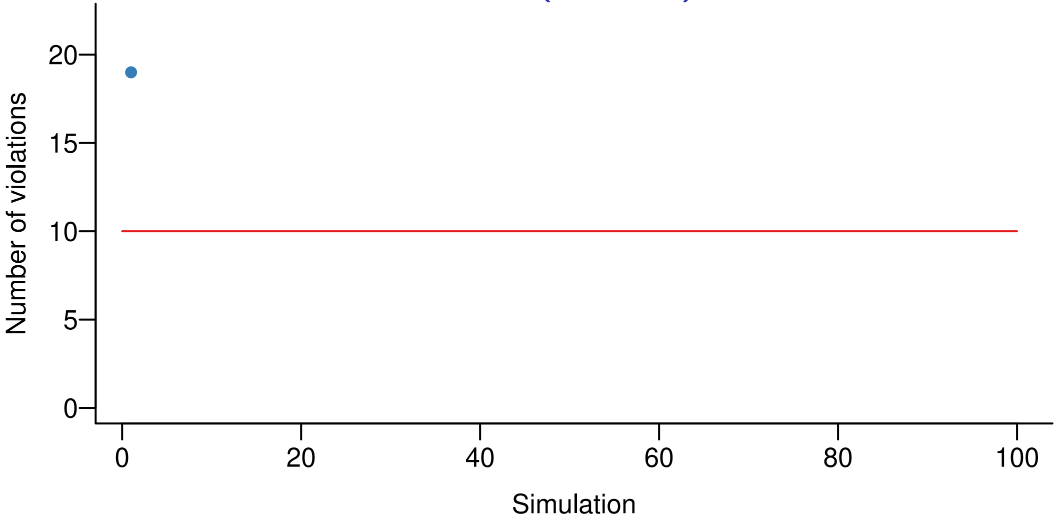
Outcome (Part I)



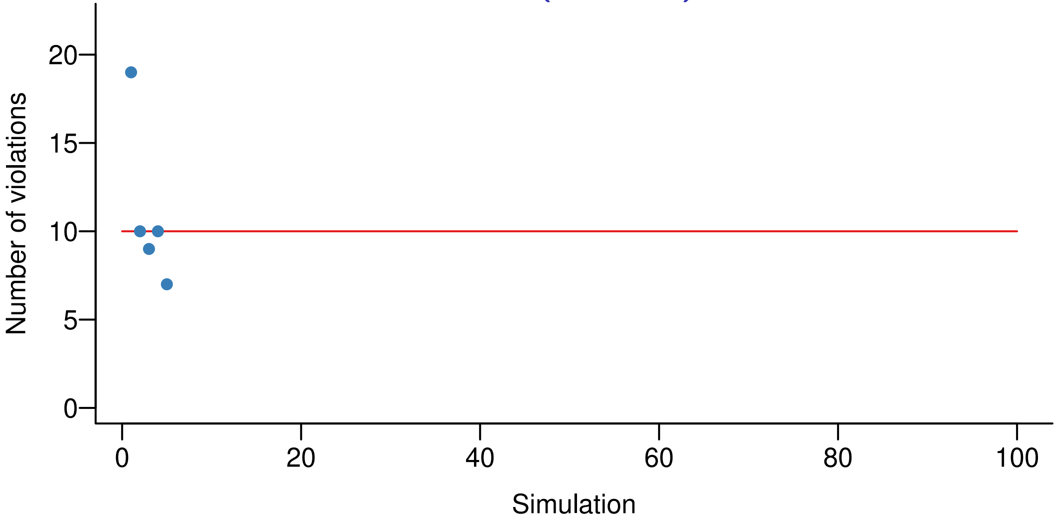
Outcome (Part II)

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

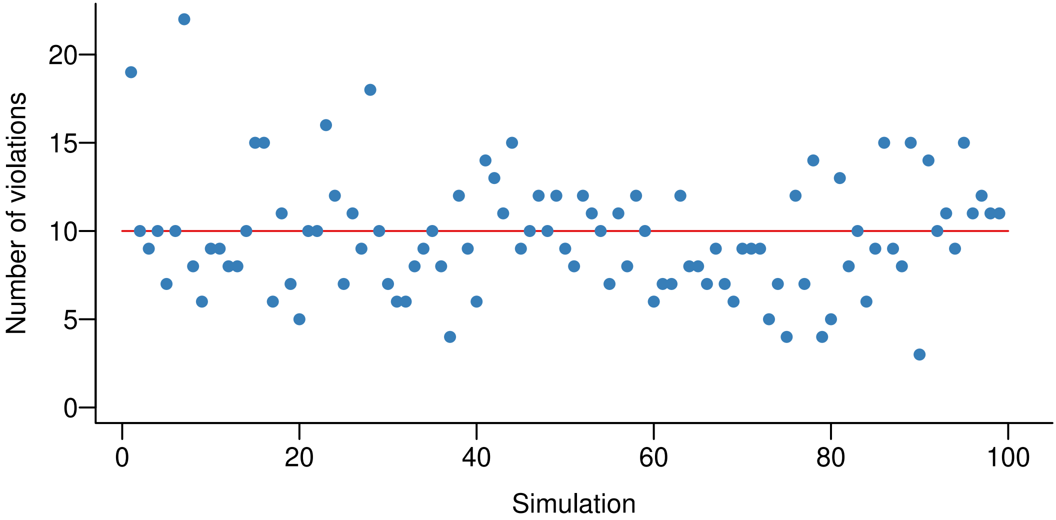
Outcome (Part III)



Outcome (Part IV)



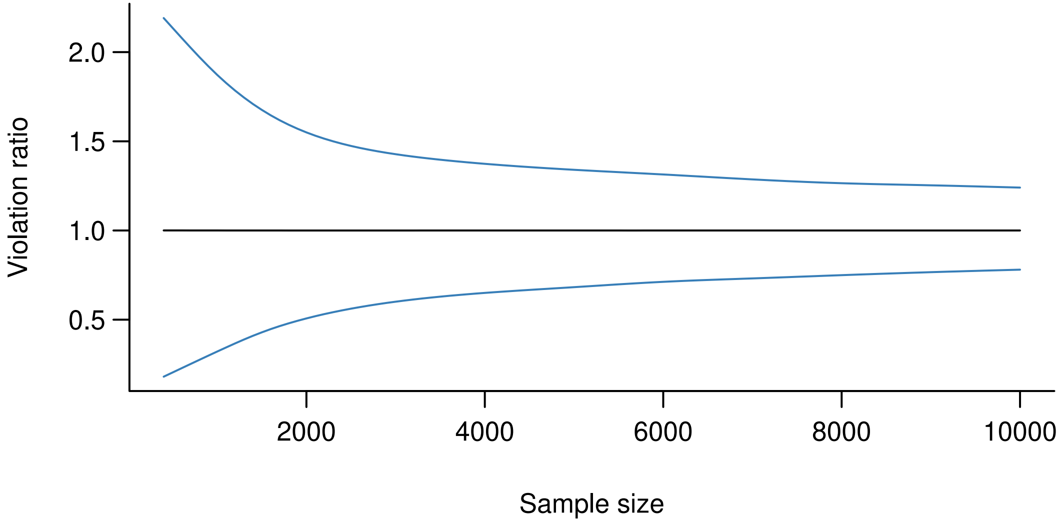
Outcome (Part V)



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



So

- Simulating many samples gives us a benchmark for expected variation
- These bounds help assess whether a model's violation ratio is significantly off
- A VR outside the 99% band suggests model failure

Application of Backtesting

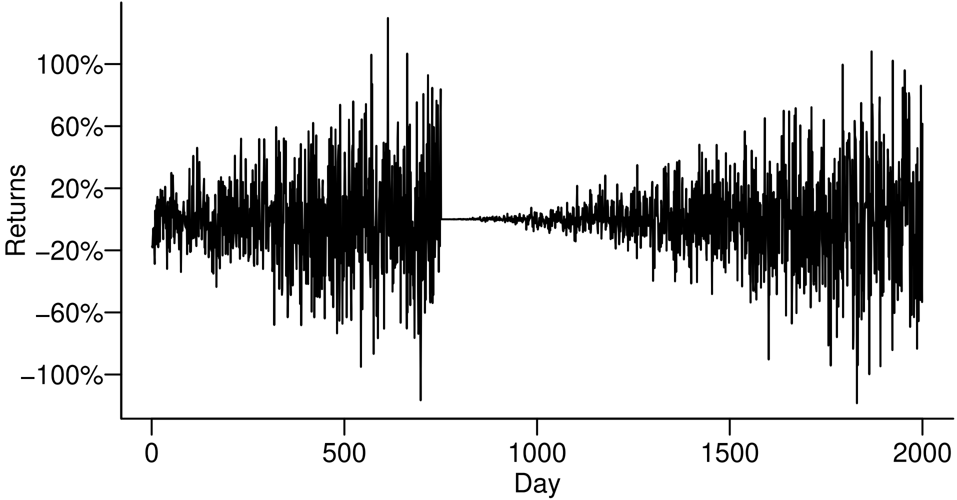
How Do VaR Models React to Volatility Regimes?

- We compare different VaR models under extreme volatility shifts
- Start with synthetic data that abruptly transitions from high to low volatility
- Then at end of Chapter examine recent real-world crises: 2008, Covid, the Ukraine war and Trump tarrifs
- Goal: observe responsiveness, smoothness and failure patterns

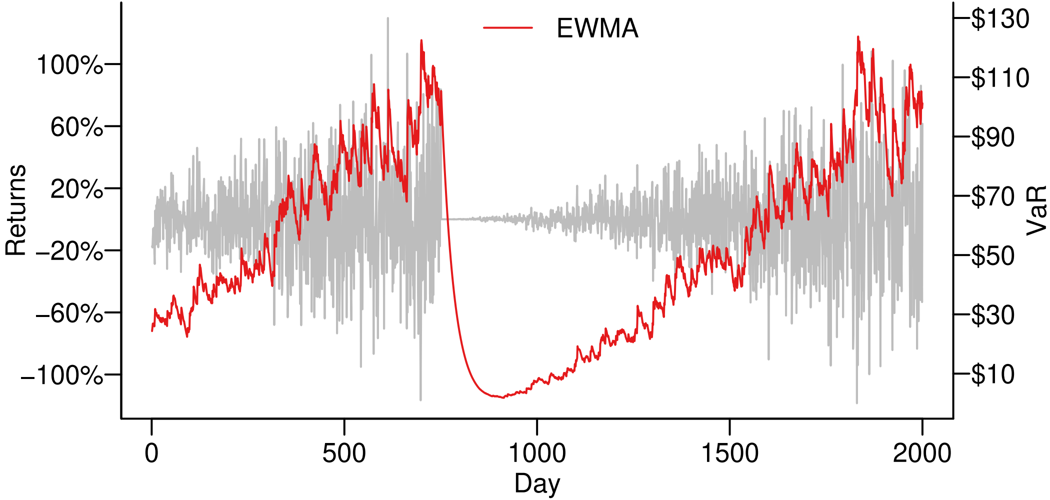
Extreme Example

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

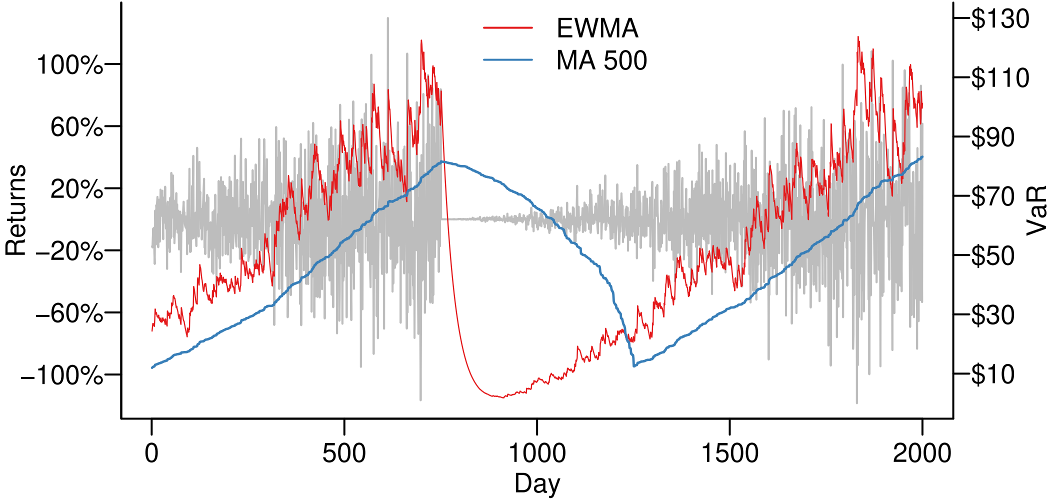
Volatility and VaR: Extreme Example



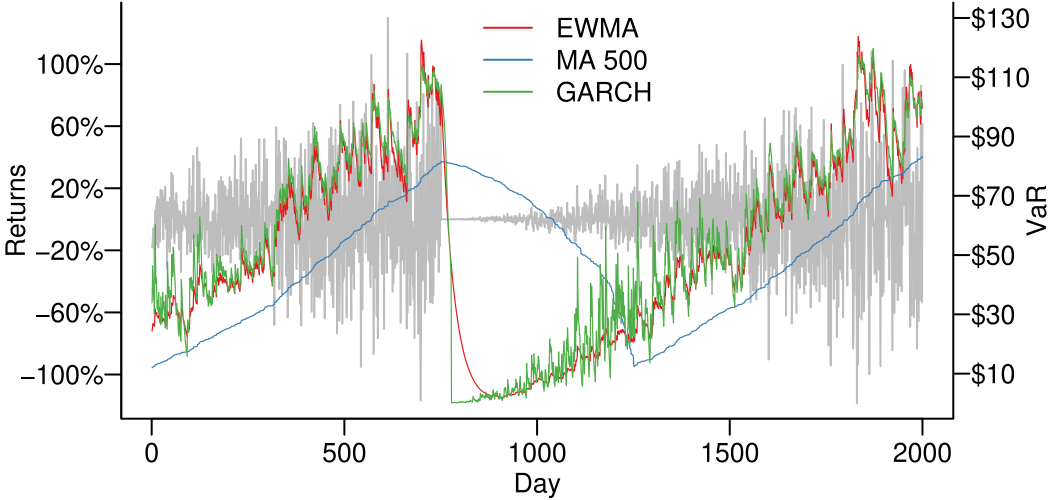
Volatility and VaR: Extreme Example



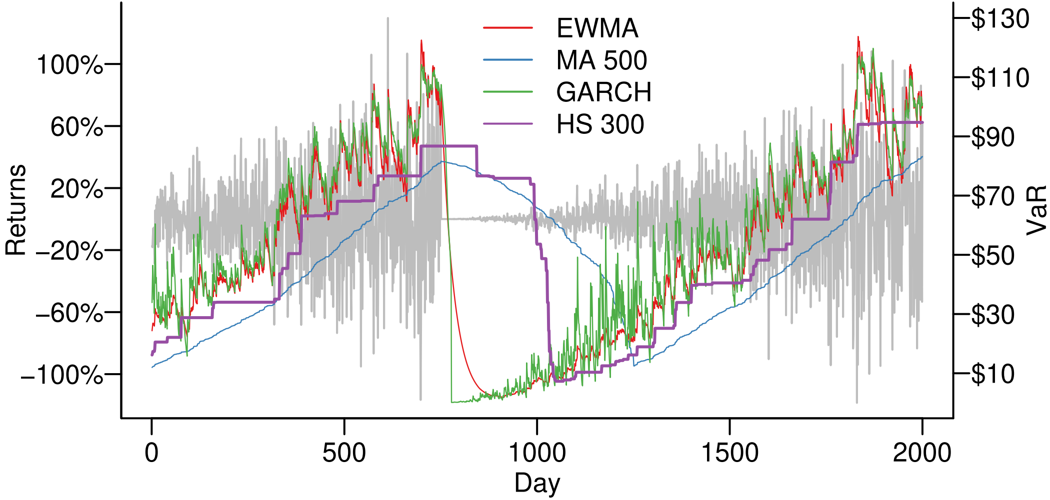
Volatility and VaR: Extreme Example



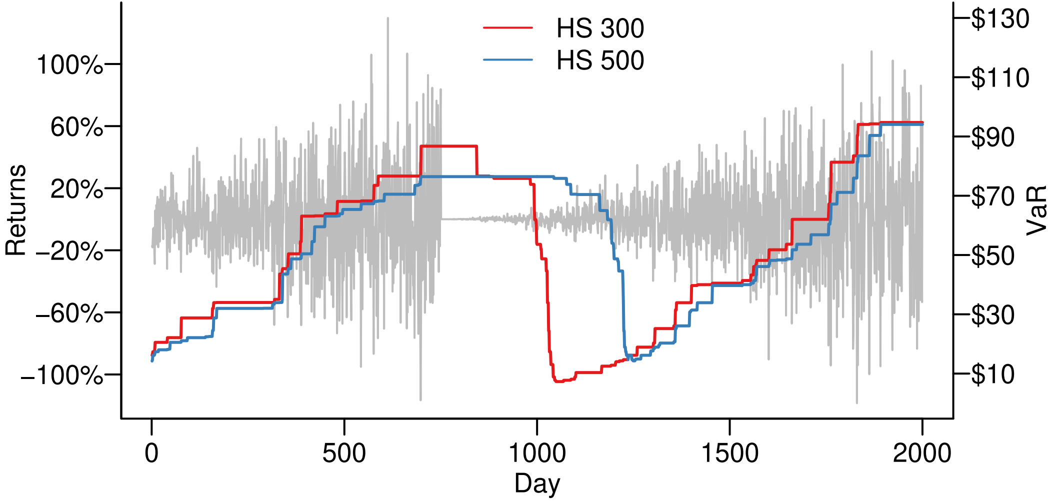
Volatility and VaR: Extreme Example



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

Significance of Backtests

Peter Christoffersen, 1998, "Evaluating Interval Forecasts" International Economic Review, 39, 841-862.

Three Backtesting Principles

- Unconditional Coverage: Do we get the right number of violations?
- Independence: Are violations randomly scattered over time?
- Joint Test: Do both hold at once?

  These are formalised in the Bernoulli testing framework by Christoffersen (1998)
- These tests are model-free and work regardless of how VaR forecasts were generated

Testing Violations

- We can test whether we get the expected number of violations and if there are patterns in the violations:
  1. The number of violations (tested by the unconditional coverage)
  2. Clustering (tested by independence tests)
- If the model is correctly specified, η_t should be a Bernoulli random variable with success probability ρ (the VaR level)

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

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

Estimation

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

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

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

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

Bernoulli coverage test

Unconditional Coverage

- Does the expected number of violations, as given by ρ match the observed number of violations from $\hat{\rho}$?
  - 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
- However, unconditional coverage alone is insufficient when violations are clustered in time

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

Likelihood

- Recall from Chapter 2 that the likelihood function is the product of the time t densities

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

- The unrestricted (using estimated probabilities, $\hat{\rho}$) likelihood function is therefore given by:

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

- Which simplifies to

$$L_U(\hat{\rho}) = (1-\hat{\rho})^{\upsilon_0}(\hat{\rho})^{\upsilon_1}$$

- Recall υ counts violations/no violations

- 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-p)*(lv-sv)
  bl=log(sv/lv)*sv +log(1-sv/lv)*(lv-sv)
  return(-2*(al-bl))
 }
```

Floating Point Numbers in Practice

- Computers use the IEEE 754 standard for representing real numbers
- Bit is 1 or 0
- A double-precision number (64-bit float) is stored as:
  - 1 bit for the sign
  - 11 bits for the exponent (range: roughly -10^{308} to 10^{308})
  - 52 bits for the fractional part (mantissa)
- This gives:
  - About 15–17 significant decimal digits of precision
  - But rounding errors for very large or very small numbers
  - Catastrophic cancellation when subtracting similar quantities
- Log-likelihoods involve multiplying (or dividing) many small probabilities — easy to lose precision

Numerical Considerations

- Note

$$(\log \mathcal{L}_U(\hat{\rho}) - \log \mathcal{L}_R(\rho)) = \log \frac{(1-\hat{\rho})^{\upsilon_0}(\hat{\rho})^{\upsilon_1}}{(1-\rho)^{\upsilon_0}(\rho)^{\upsilon_1}}$$

- but

```
log(sv/lv)*sv + log(1-sv/lv)*(lv-sv)
- log(p)*sv - log(1-p)*(lv-sv)
```

- *Can be* different from

```
log((sv/lv)^sv * (1-sv/lv)^(lv-sv)/
          (p^sv * (1-p)^(lv-sv)))
```

- Where the latter is more likely to lose precision

Independence Property

Distribution of Violations

- If violations tend to appear in clusters, something is wrong
- A good risk model should adapt quickly after a breach — not allow repeated surprises
- The indicator $\eta_t = 1$ (violation), $\eta_t = 0$ (no violation) should look random
- 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
- We test this using a first-order Markov chain that only considers yesterday's violation state

Setup

- Look at violations today and yesterday

$$\eta_{\text{yesterday,today}}$$

- There are 4 cases

η_{00} No violation yesterday, no violation today

η_{11} Violation yesterday, violation today

η_{01} No violation yesterday, violation today

η_{10} Violation yesterday, no violation today

- The theoretic probabilities are

$$\rho_{ij}$$

- The count of each case is

$$v_{ij}$$

- Total number of observations

$$v_{00} + v_{10} + v_{01} + v_{11}$$

What Is a Markov Chain?

- A Markov chain is a stochastic process where the future depends only on the present state — not the past history
- It is defined by a set of states and transition probabilities between them
- In our case:
  - The state is whether a VaR violation occurred ($\eta_t = 1$) or not ($\eta_t = 0$)
  - We model the probability of a violation today based on whether there was one yesterday
- If the VaR model is correct, transitions should be independent:

$$\mathbb{P}(\eta_t = 1 \mid \eta_{t-1} = 1) = \mathbb{P}(\eta_t = 1 \mid \eta_{t-1} = 0)$$

- This leads us to test whether observed transitions differ from this independence

Markov Transition Matrix for Violations

- Define a first-order Markov chain where today's violation depends on yesterday's
- We estimate four frequencies: $\eta_{00}, \eta_{01}, \eta_{10}, \eta_{11}$
- The first order *transition probability matrix* is defined as

	$\eta_t = 0$	$\eta_t = 1$	sum
$\eta_{t-1} = 0$	ρ_{00}	ρ_{01}	$\rho_{00} + \rho_{01}$
$\eta_{t-1} = 1$	ρ_{10}	ρ_{11}	$\rho_{10} + \rho_{11}$
sum	$\rho_{00} + \rho_{10}$	$\rho_{01} + \rho_{11}$	1

Towards the null hypothesis

- Under the null hypothesis of no clustering, the probability of a violation tomorrow *does not depend* on today being a violation
- Then

$$\rho_{01} = \rho_{11} := \rho$$

- And

$$\rho_{10} = \rho_{00} = 1 - \rho$$

- So the null transition matrix is simply:

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

Estimating the transition matrix

- Use the estimated number of violations

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

- Then we want to test if

$$\hat{\Pi} = \Pi$$

Likelihood Ratio Test

- The likelihood under the null (constrained) is same as in the coverage test

$$L(\Pi) = (1-\rho)^{v_{00}+v_{10}}\,\rho^{v_{01}+v_{11}}$$

- The approximate unconstrained likelihood function, because there are two states, is the product of each state:

$$L(\hat{\Pi}) = (1-\hat{\rho}_{01})^{v_{00}}\,\hat{\rho}_{01}^{v_{01}}\,(1-\hat{\rho}_{11})^{v_{10}}\,\hat{\rho}_{11}^{v_{11}}$$

- The LR test is then:

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

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
- The test only detects first-order dependence (one-day memory)
- 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

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

What Does the Independence Test Tell Us?

- Tests whether VaR violations follow a memoryless process
- Detects models that react too slowly to changing risk
- Joint test with coverage provides stronger model validation
- We now apply it to S&P 500 models

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

So

- MA is rejected strongly — both coverage and independence
- GARCH fails coverage but passes independence
- HS passes in early sample, fails in full sample
- EWMA performs reasonably — especially in early data

Expected Shortfall Backtesting

Why Expected Shortfall is Hard to Backtest

- Expected Shortfall (ES) is the average loss in the worst $\rho\%$ of cases:

$$\text{ES} = -\text{E}[q \mid q > \text{VaR}]$$

- But ES is not defined by a single threshold breach — it depends on the full shape of the tail
- Unlike VaR, ES is:
  - Not directly linked to a simple event (like a violation)
  - Not *elicitable* on its own (you cannot score forecasts with a simple loss function)
  - Sensitive to rare, extreme losses — which occur infrequently
- Result: Traditional frequency-based backtesting (counting violations) does not apply
- ES backtests are necessarily approximated and unavoidably sensitive to VaR prediction errors

Joint Elicitability of VaR and ES

- Although ES is not elicitable alone, it is jointly elicitable with VaR
- This was shown by Fissler and Ziegel (2016)
- Joint scoring functions allow consistent evaluation of forecasts
- Enables meaningful comparison and backtesting of models producing both VaR and ES
- All ES backtests (including Acerbi-Székely) rely on this joint property
- Consequence: ES backtest accuracy depends on VaR prediction quality

The Acerbi-Székely (2019) ES Backtest

- Since ES cannot be tested directly via violation counting:
  - Define a scoring function that penalises ES forecast errors
  - Evaluate whether observed losses align with forecast ES given VaR violations
- Key insight: The test function

$$Z_{\mathrm{ES}}(\mathrm{ES}_t, \mathrm{VaR}_t, q_t) = \mathrm{ES}_t - \mathrm{VaR}_t - \frac{1}{\rho}(q_t + \mathrm{VaR}_t)_+$$

  compares forecast ES with VaR plus the average excess loss beyond VaR
- The term $\frac{1}{\rho}(q_t + \mathrm{VaR}_t)_+$ represents the mean tail loss when losses exceed VaR
- Here $(x)_+ = \max(x, 0)$ is the positive part function
- Under perfect forecasting, this should equal the difference between ES and VaR

Bias in the Acerbi-Székely Test

- Expected value: $\mathrm{E}[Z_{\mathrm{ES}}] = \mathrm{ES}_t - \mathrm{ES}_t^{\mathrm{true}} - B(\mathrm{VaR}_t)$ where $B(\mathrm{VaR}_t) \geq 0$ is a bias term
- This bias is:
  - Zero when VaR predictions are perfect ($\mathrm{VaR}_t = \mathrm{VaR}_t^{\mathrm{true}}$)
  - Small when VaR predictions are reasonably accurate
  - Prudential: makes the test more conservative, not more lenient

Acerbi-Székely Test: Implementation

- Define the scaled score at time t using realised loss q_t, forecast VaR and ES:

$$S_t = \frac{q_t - \mathrm{ES}_t}{\mathrm{VaR}_t - \mathrm{ES}_t} \cdot \mathbb{1}_{\{q_t > \mathrm{VaR}_t\}}$$

- Here $\mathbb{1}_{\{q_t > \mathrm{VaR}_t\}}$ is the indicator function: equals 1 if $q_t > \mathrm{VaR}_t$, zero otherwise
- Under correct forecasts, the sequence $\{S_t\}$ should have mean zero and finite variance

Why the Test Statistic Follows $\mathcal{N}(0,1)$

- The test statistic is:

$$Z = \frac{\sqrt{W_T} \cdot \bar{S}}{\hat{\sigma}}$$

  where $\bar{S} = \frac{1}{W_T} \sum_{t=1}^{W_T} S_t$
- Under the null hypothesis (correct forecasts):
  - Sample mean $\bar{S}$ has expected value zero
  - Central Limit Theorem: $\bar{S}$ is approximately normal with variance σ^2/W_T
  - Standardisation: dividing by estimated standard deviation $\hat{\sigma}$ gives unit variance
- Result: $Z \sim \mathcal{N}(0,1)$ under null hypothesis of correct ES and reasonably accurate VaR

Bias Properties and Practical Implications

- The bias term $B(\text{VaR}_t)$ in the Acerbi-Székely test is:
  - Quadratic in small VaR discrepancies: $B(\text{VaR}_t) \approx \frac{f(-\text{VaR}_t^{\text{true}})}{2\rho}(\text{VaR}_t - \text{VaR}_t^{\text{true}})^2$
  - Here f is the pdf of the actual return distribution (not necessarily normal)
  - Typically negligible when VaR predictions are within $\pm 15\%$ of true values
  - Still manageable when VaR accuracy is within $\pm 40\%$
- Prudential nature means:
  - Imperfect VaR predictions make ES tests more stringent, not more lenient
  - Type II errors (accepting bad models) are less likely
  - Conservative approach suitable for risk management
- Practical requirement: VaR models should be reasonably well-calibrated before ES backtesting

Limitations and Practical Considerations

- The reliability of any ES backtest procedure is lower than that of VaR
  - With ES, we test whether the mean of returns on days when VaR is violated equals the average ES forecasts on these days
  - Much harder to create formal tests than the coverage tests for VaR violations
  - Test accuracy depends critically on VaR prediction quality
- ES backtesting requires many more observations than VaR backtesting
- The Acerbi-Székely approach requires VaR predictions to be reasonably accurate
- When ES is obtained directly from VaR and provides the same signal as VaR (when VaR is subadditive), VaR backtesting may be more reliable

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

Types of Stress Tests

- **Sensitivity** — shock one risk factor at a time ($+100$ bp rates, -10% equity)
- **Historical** — replay a past crisis window (2008–09, 2020 Q1)
- **Hypothetical** — forward-looking narrative combining multiple shocks (pandemic plus stagflation)
- **Reverse** — search for the smallest scenario that breaches a capital or liquidity limit

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	2010-2015
Brexit	2017
Covid-19	2020

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

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

Recent Stress Events

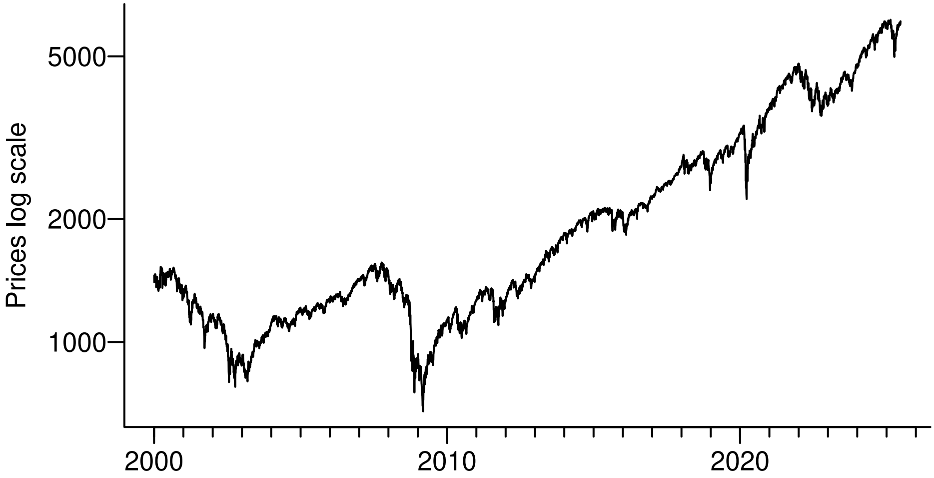
Backtesting the S&P-500 in Times of Stress

- Make the estimation window 1,000 days
- Probability: 1%
- Portfolio value 1,000
- And compare GARCH, tGARCH and historical simulation

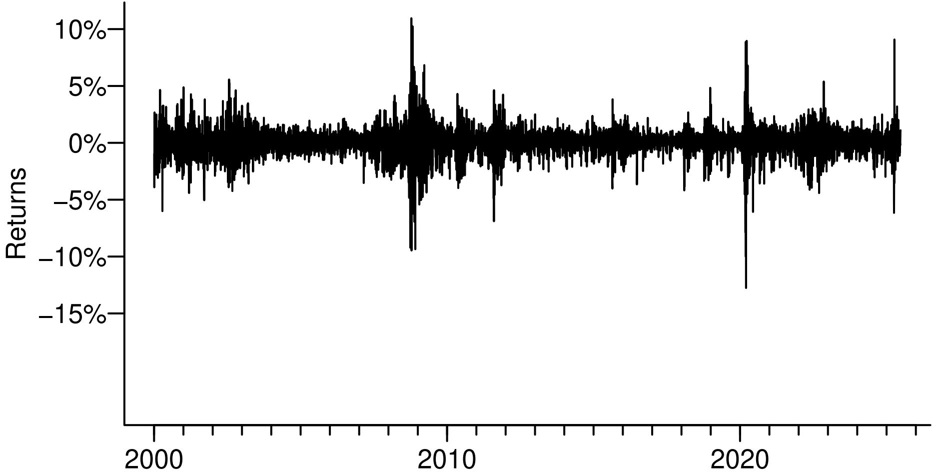
S&P-500

- We then take a sample from the S&P-500
- There are some especially interesting regimes,
- The collapse of volatility after 2003 and the crisis in 2008
- Covid in 2020
- Russia-Ukraine war and inflation in 2022-2024
- Trump tarrifs 2025

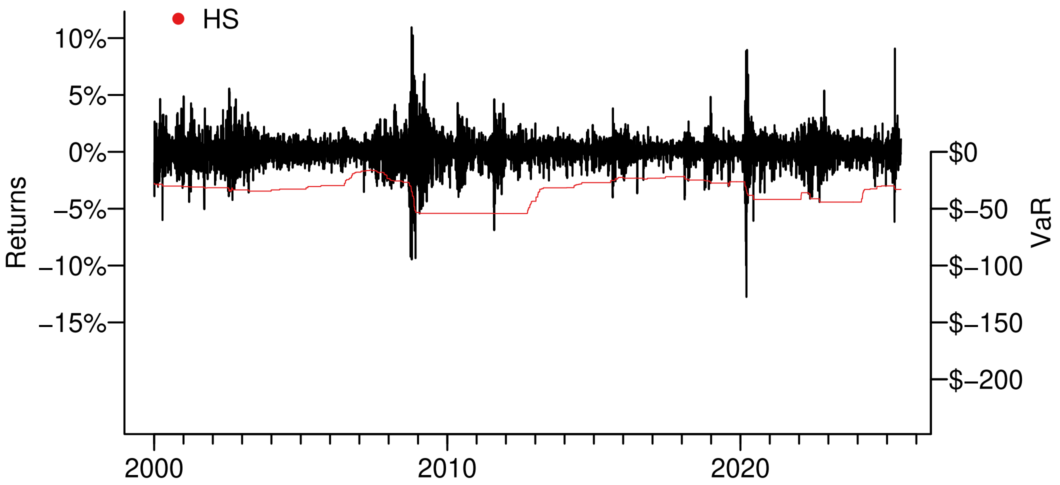
S&P-500 2000 to June 2025



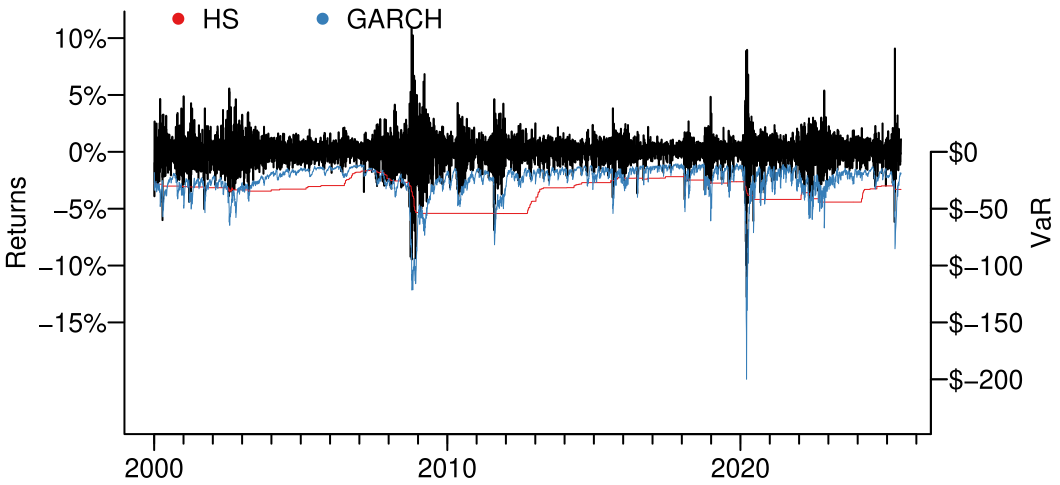
S&P-500 2000 to June 2025



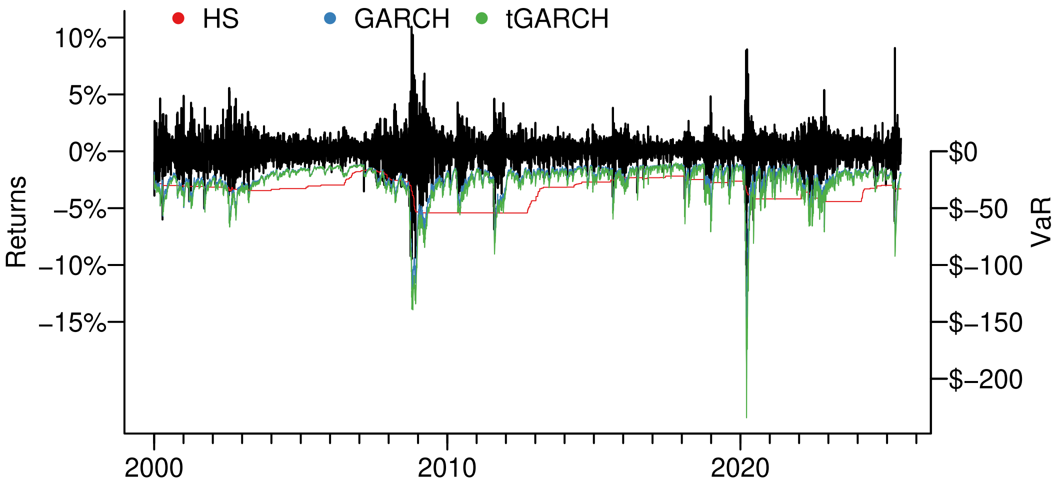
S&P-500 2000 to June 2025



S&P-500 2000 to June 2025



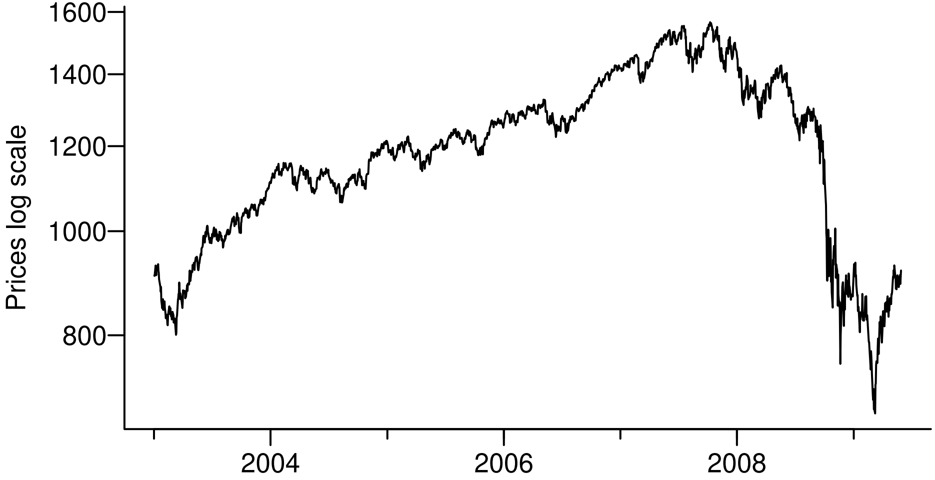
S&P-500 2000 to June 2025



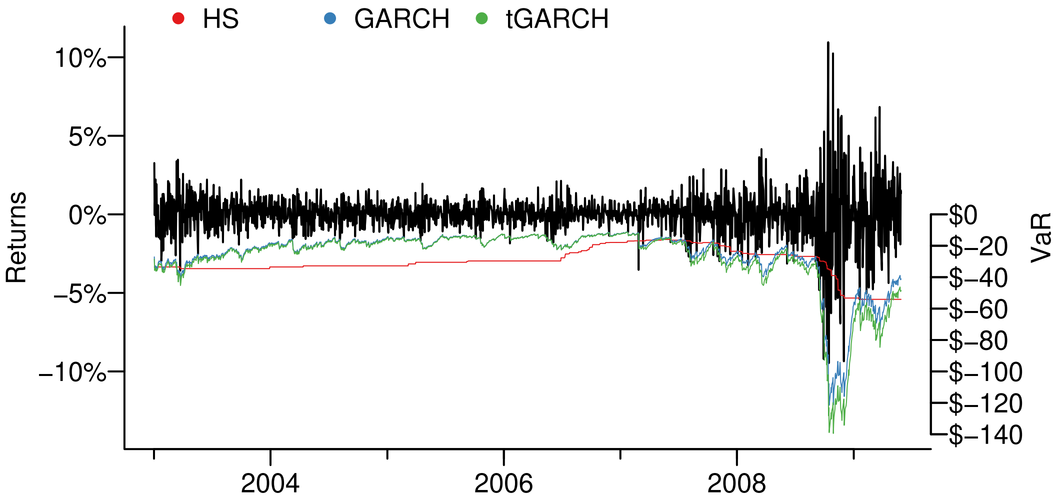
(Almost) Global Crisis in 2008

- “Great Moderation” 2003-2007
- Followed by a crisis that hit many countries
- Started in June 2007 with a quant fund crisis
- Investors “went on strike” in July 2007
- Crisis peaked in end of September 2007/October 2008
- Intensive crisis phase over early December
- Markets begin to recover in early 2009

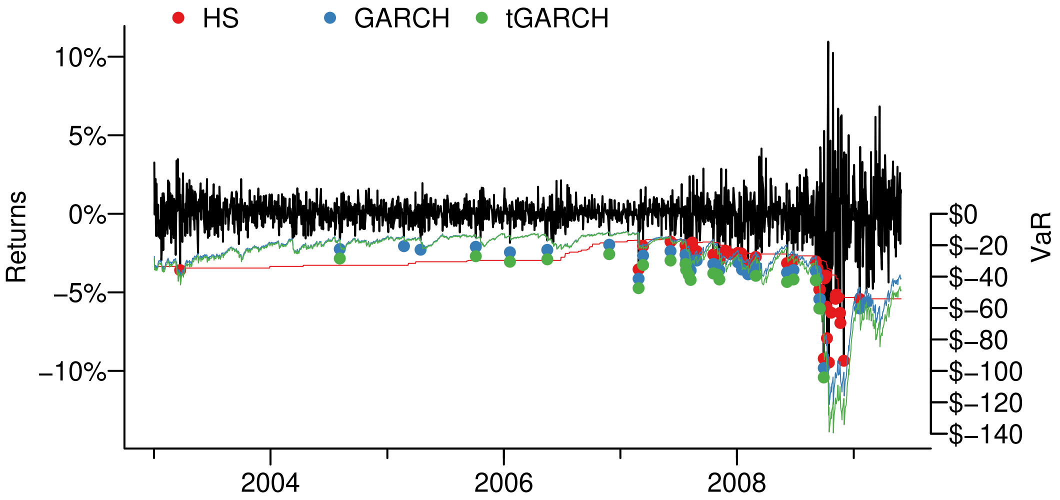
(Almost) Global Crisis in 2008



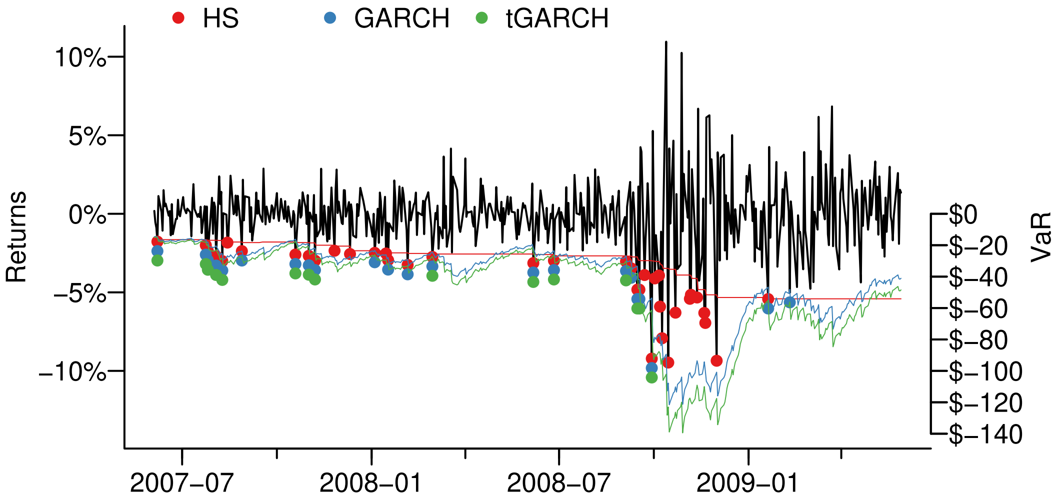
(Almost) Global Crisis in 2008



(Almost) Global Crisis in 2008



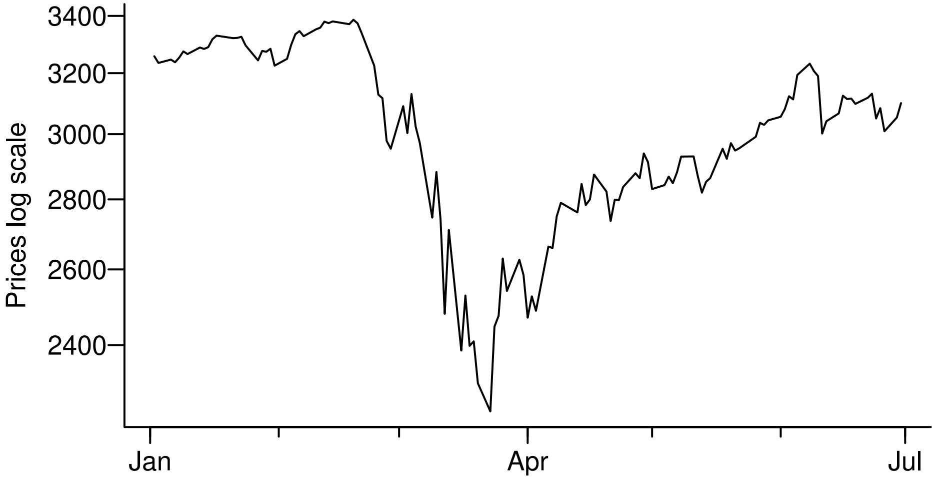
(Almost) Global Crisis in 2008 — Zoom into Crisis



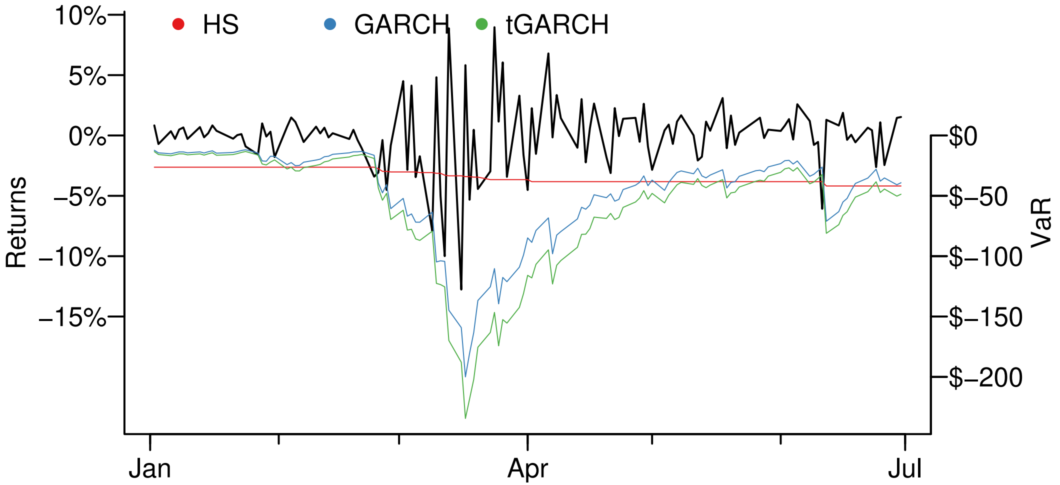
Covid-19 2020

- Markets in China fall in February
- Rest of world starts in March
- 14 April worst day
- Market recover quickly
- "V" shape crisis

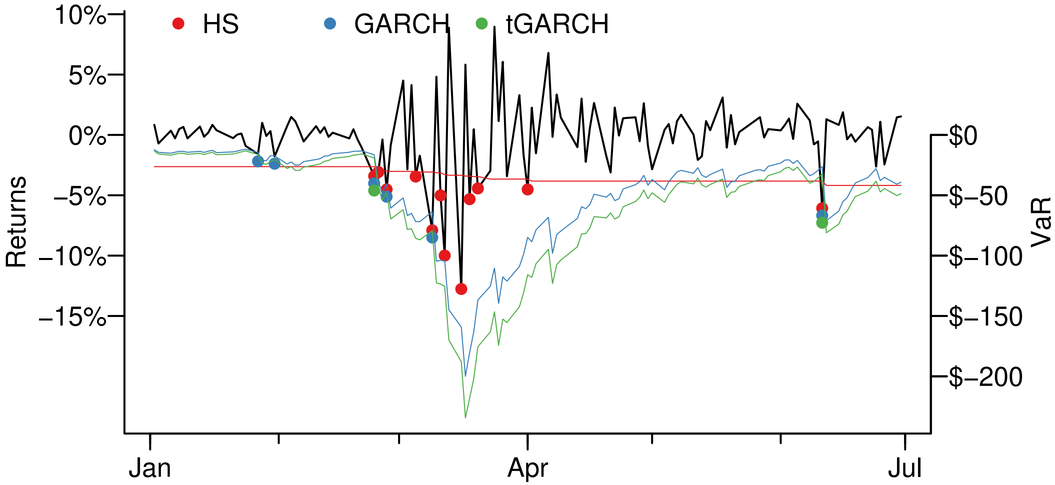
Covid-19 2020



Covid-19 2020



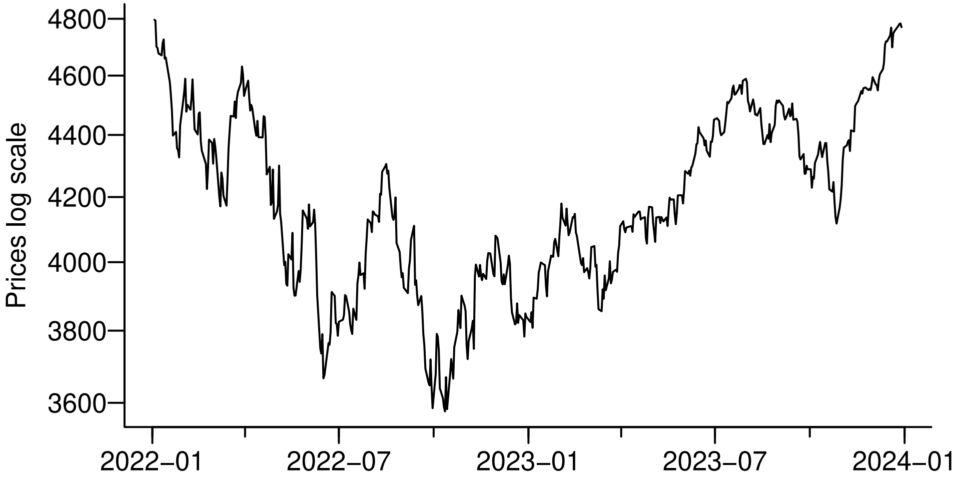
Covid-19 2020



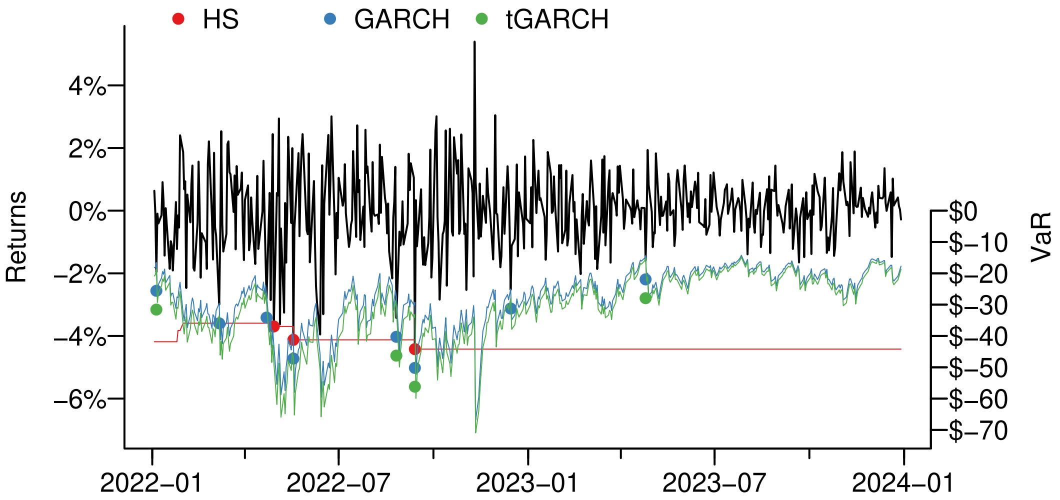
2022-2024

- Russia invades Ukraine
- Very little impact on US (and hence S&P-500)
- Biggest impact on Germany
- The inflation shock that year and next more important

2022-2024


4800
4600
4400
4200
4000
3800
3600
Prices log scale
2022–01
2022–07
2023–01
2023–07
2024–01

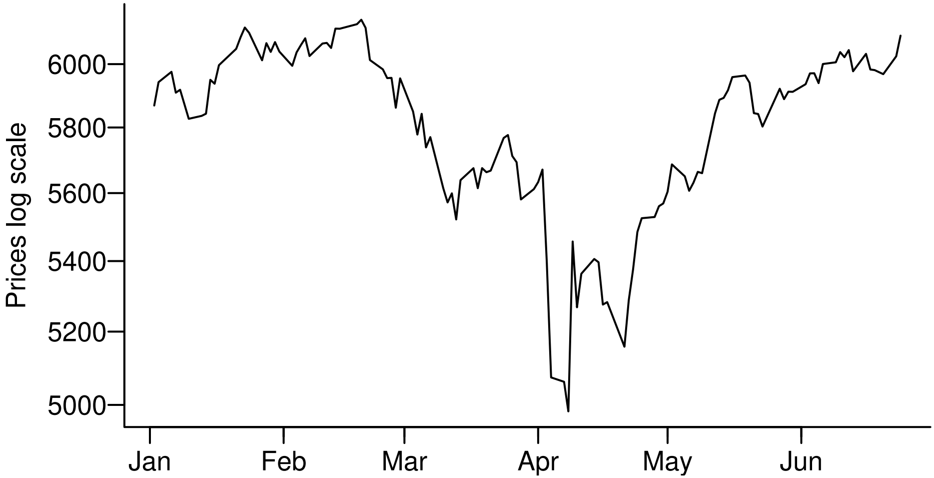
2022-2024


HS
GARCH
tGARCH
Returns
4%
2%
0%
−2%
−4%
−6%
VaR
$0
$−10
$−20
$−30
$−40
$−50
$−60
$−70
2022−01
2022−07
2023−01
2023−07
2024−01

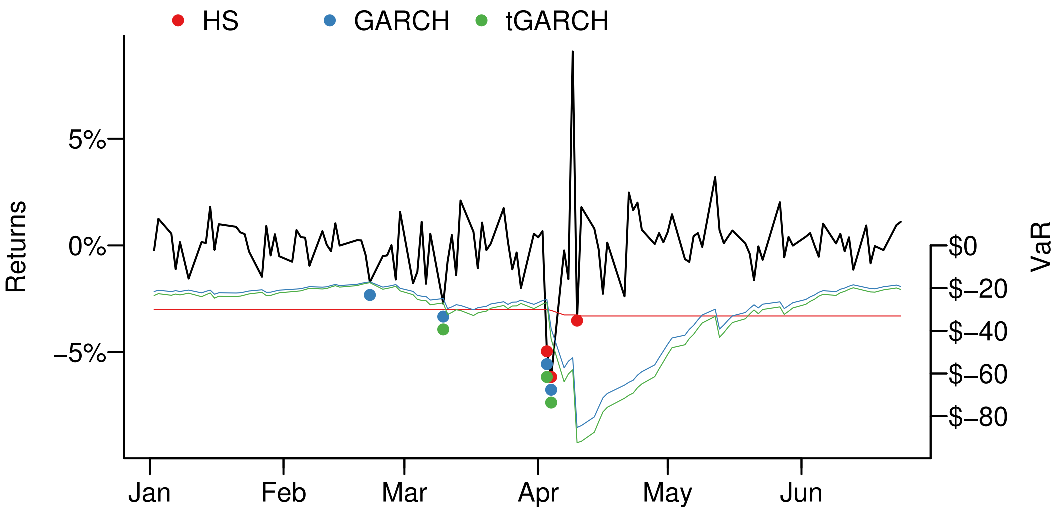
Trump Tarrifs

- Initial market reaction to Trump positive
- Shock on announcement dates
- Very quick recovery
- What do you think these results say about the market consensus on the impact of Trump?

2025


Prices log scale
6000
5800
5600
5400
5200
5000
Jan
Feb
Mar
Apr
May
Jun

2025


HS
GARCH
tGARCH
5%
0%
−5%
Returns
$0
$−20
$−40
$−60
$−80
VaR
Jan
Feb
Mar
Apr
May
Jun

What We Learn from Backtesting in Crisis

- VaR models differ in responsiveness to regime shifts
- GARCH can overshoot after volatility collapses
- Real-world backtesting reveals limits of statistical calibration
- Visual inspection complements formal testing

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 may not be violated on the worst day of the crisis
- Why do you think that is the case?
- These results confirm what we have found for the same methods in other cases