Actuarial Research Centre

Institute and Faculty of Actuaries

Actuarial Research Centre (ARC)

PhD studentship output

The Actuarial Research Centre (ARC) is the Institute and Faculty of Actuaries' network of actuarial researchers around the world. The ARC seeks to deliver research programmes that bridge academic rigour with practitioner needs by working collaboratively with academics, industry and other actuarial bodies.
The ARC supports actuarial researchers around the world in the delivery of cutting-edge research programmes that aim to address some of the significant challenges in actuarial science.

Multinomial VaR Backtests

Yen H. Lok

Acknowledgement:
Alexander J. McNeil, Michael Gordy and Marie Kratz.

Heriot Watt University
12 Jan 2017

Overview

Fundamental Review of the Trading Book (FRTB)

In order to be approved for an internal-model approach, banks is required to:

- Conduct regular backtesting and P&L attribution programmes.
- Backtesting requirements are based on comparing each desk's 1-day static value-at-risk measure (using the most recent 12 months' data) at both the 97.5th and 99th percentile to the daily trading outcome.

Additionally, banks may be required to:

- Testing carried out for longer periods than required for the regular backtesting programme (eg three years); or
- Testing carried out using the entire forecasting distribution. For example the bank could be required to use report the following:
  - (i) A daily ES calibrated to 97.5th level;
  - (ii) The daily P&L for the desk; and
  - (iii) The *p*-value for the P&L on each day for each desk.

Value-at-Risk (VaR)

- The Value at Risk (VaR) is defined as the generalized inverse of the forecast model F, given by

$$\text{VaR}_\alpha := F^{\leftarrow}(\alpha) = \inf\{l \in \mathbb{R} : F(l) > \alpha\}.$$

- When the forecast distribution is continuous, the VaR is simply the ordinary inverse of the forecast distribution.

Spectral risk measures and expected shortfall

- A Spectral Risk Measure $\mathcal{M}_g$ with an admissible risk spectrum g is defined as

$$\mathcal{M}_g := \int_0^1 g(u)\,\mathrm{VaR}_u\,\mathrm{d}u.$$

- Spectral Risk Measures are weighted integrals of the VaR, where the weight function g is required to satisfy certain constraints.
- We say that g is an admissible risk spectrum if
  - i g is non-negative
  - ii $\int_0^1 g(u)\mathrm{d}u = 1$
  - iii g is non-decreasing
- The Expected Shortfall, ES_α, is defined as

$$\mathrm{ES}_\alpha := \frac{1}{1-\alpha}\int_\alpha^1 \mathrm{VaR}_u\,\mathrm{d}u,$$

is a special case of Spectral Risk Measure, with $g(u) = \frac{1}{1-\alpha} I_{\{\alpha \le u \le 1\}}$, and share the properties of Spectral Risk Measures.

Binomial score test

- We denote L_t as the realized loss at time t, and refer to the event $\{L_t > \text{VaR}_\alpha\}$ as a VaR exception at level α.
- We define the exception process at level α to be the process $I_{t,\alpha} := I_{\{L_t > \text{VaR}_\alpha\}}$ for $t = 1, \ldots, n$.
- When the losses L_t have distribution F, assuming that F is continuous, it is well known (Christoffersen, 1998) that the sequence $(I_{t,\alpha})_{t=1,\ldots,n}$ should satisfy:
  - the unconditional coverage hypothesis, $E(I_{t,\alpha}) = 1 - \alpha$ for $\forall t$, and
  - the independence hypothesis, $I_{t,\alpha}$ is independent of $I_{s,\alpha}$ for $s \neq t$.
- A test for the unconditional coverage hypothesis is the binomial score test

$$Z_\alpha := \frac{\sum_{t=1}^{n} I_{t,\alpha} - n(1-\alpha)}{\sqrt{n\alpha_1(1-\alpha)}},$$

which is compared with a standard normal distribution.

Testing VaR at multiple levels

- For a series of $\text{VaR}_{\boldsymbol{\alpha}}$, at a series of ordered levels $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)$, with $\alpha_0 = 0$ and $\alpha_{N+1} = 1$, we define the exception indicator at the level α_j at time t by

$$I_{t,\alpha_j} := I_{\{L_t > \text{VaR}_{\alpha_j}\}}.$$

- We define $X_t = \sum_{j=1}^{N} I_{t,\alpha_j}$ which counts the number of VaR levels that are breached. The sequence (X_t) should satisfy:
  - the unconditional coverage hypothesis, $\text{P}(X_t \leq j) = \alpha_{j+1},\ j = 0, \ldots, N$ for $\forall t$, and
  - the independence hypothesis, X_t is independent of X_s for $s \neq t$.
- We now define observed cell counts by

$$O_j = \sum_{t=1}^{n} I_{\{X_t = j\}}, \quad j = 0, 1 \ldots, N,$$

then the random vector $(O_0, \ldots, O_N)$ should follow the multinomial distribution

$$(O_0, \ldots, O_N) \sim \text{MN}(n, (\alpha_1 - \alpha_0, \ldots, \alpha_{N+1} - \alpha_N)) \ .$$

Pearson chi-squared and Nass test

- A well known test for the multinomial distribution is the Pearson chi-squared test

$$S_N = \sum_{j=0}^{N} \frac{(O_{j+1} - n(\alpha_{j+1} - \alpha_j))^2}{n(\alpha_{j+1} - \alpha_j)},$$

  where under the null hypothesis, S_N is asymptotically χ^2_N distributed.
- It is well known that the accuracy of this test decreases with increasing N.
- Nass (1959) studied an improved approximation to the distribution of the statistic S_N, using the r.v. cS_N, where

$$c\,S_N \sim \chi^2_\nu\,, \quad \text{with} \quad E(c\,S_N) = v \ \text{and} \ var(c\,S_N) = 2\nu.$$

- Pearson (1932) has shown that

$$E(S_N) = N\,, \quad \text{and} \quad var(S_N) = 2N - \frac{N^2 + 4N + 1}{n} + \frac{1}{n}\sum_{j=0}^{N} \frac{1}{\alpha_{j+1} - \alpha_j}.$$

- The Nass test offers an appreciable improvement over the chi-square test when cell probabilities are small.

Probit Normal LRT

- The LRT statistic for a general multinomial test is given by

$$\tilde{S}_N = 2\sum_{j=0}^{N} O_j \ln\left(\frac{\hat{\theta}_{j+1} - \hat{\theta}_j}{\alpha_{j+1} - \alpha_j}\right),$$

  where $\hat{\theta}_{j+1} - \hat{\theta}_j = O_j/n$.
- When O_j is zero for some j, the test statistic is undefined.
- We consider a general model in which the parameters are given by

$$\theta_j = \Phi\left(\frac{\Phi^{-1}(\alpha_j) - \mu}{\sigma}\right), \quad j = 1, \ldots, N,$$

  where $\mu \in \mathbb{R}$, $\sigma > 0$ and Φ denotes the standard normal distribution function. We test the null hypothesis $H_0 : \mu = 0$ and $\sigma = 1$ against the alternative $H_1 : \mu \neq 0$ or $\sigma \neq 1$.
-

$$\hat{\theta}_{j+1} - \hat{\theta}_j = \Phi\left(\frac{\Phi^{-1}(\alpha_{j+1}) - \hat{\mu}}{\hat{\sigma}}\right) - \Phi\left(\frac{\Phi^{-1}(\alpha_j) - \hat{\mu}}{\hat{\sigma}}\right),$$

  where $\hat{\mu}$ and $\hat{\sigma}$ are the MLEs under H_1, and the test statistic $\tilde{S}_N$ is asymptotically χ^2_2 distributed.

Experimental design

- In each experiment we generate a total dataset of n values from the true distribution G.
- We consider the cases when G is normal, Student distributions with 5 and 3 degrees of freedom ($t5$ and $t3$) which have moderately heavy and heavy tails respectively, and the skewed Student distribution of Fernandez & Steel (1998) with 3 degrees of freedom and a skewness parameter $\gamma = 1.2$ (denoted $st3$). We normalized G to have mean zero and unit variance.
- The forecast distribution F is set to be the standard normal distribution, hence no parameter estimation is required.
- The following colour coding is used: green indicates good results ($\leq 6\%$ for the size; $\geq 70\%$ for the power); red indicates poor results ($\geq 9\%$ for the size; $\leq 30\%$ for the power); dark red indicates very poor results ($\geq 12\%$ for the size; $\leq 10\%$ for the power).
- The experiment is repeated $10,000$ times to determine rejection rates.

Multinomial tests Size and power

	test	Pearson			Nass			LRT		
G	$n \mid N$	1	4	8	1	4	8	1	4	8
Normal	250	3.9	5.6	8.5	3.9	5.0	4.7	7.5	6.5	6.5
	500	3.9	5.2	6.6	3.9	4.7	4.7	5.9	5.5	5.6
	1000	5.0	5.0	5.6	5.0	4.7	4.9	4.1	5.5	5.8
t5	250	4.1	14.1	20.8	4.1	12.8	14.1	6.9	15.8	21.6
	500	5.2	22.1	28.4	5.2	20.5	24.5	6.5	26.9	36.6
	1000	6.9	40.2	48.2	6.9	39.5	46.2	5.2	46.4	61.8
t3	250	3.6	13.7	21.1	3.6	12.1	14.8	10.3	24.4	35.4
	500	4.8	25.2	32.7	4.8	22.4	28.7	9.5	44.2	58.6
	1000	9.9	55.6	62.9	9.9	54.1	60.3	9.7	75.4	87.7
st3	250	5.4	28.8	40.0	5.4	26.3	30.5	8.0	33.5	46.5
	500	6.9	50.7	60.6	6.9	47.6	56.2	7.9	59.3	73.6
	1000	9.5	83.0	89.1	9.5	82.3	88.1	6.9	88.1	95.3

Table: Estimated size and power of three different types of multinomial test (Pearson, Nass, likelihood-ratio test (LRT)) based on exceptions of N levels. Results are based on 10000 replications. $\alpha_j = 0.975 + j/N(1 - 0.975)$, $j = 0, \ldots, N$.

Binomial vs multinomial

G	n | test	Bin (0.99)	Pearson (4)	Nass (4)	LRT (4)	LRT (8)
Normal	250	4.0	5.6	5.0	6.5	6.5
	500	3.7	5.2	4.7	5.5	5.6
	1000	3.8	5.0	4.7	5.5	5.8
t5	250	17.7	14.1	12.8	15.8	21.6
	500	22.4	22.1	20.5	26.9	36.6
	1000	33.0	40.2	39.5	46.4	61.8
t3	250	13.5	13.7	12.1	24.4	35.4
	500	16.2	25.2	22.4	44.2	58.6
	1000	22.3	55.6	54.1	75.4	87.7
st3	250	31.2	28.8	26.3	33.5	46.5
	500	44.2	50.7	47.6	59.3	73.6
	1000	66.2	83.0	82.3	88.1	95.3

Table: Comparison of estimated size and power of the binomial score test with $\alpha = 0.99$ and Pearson, Nass and LRT with $N = 4$ and LRT with $N = 8$. Results are based on 10000 replications

Historical simulation model

- We now consider the industry modeller who uses the empirical distribution function by forming standard empirical quantile estimates, a method known as historical simulation in industry.
- We mimic the procedure used in practice where models are continually updated to use the latest market data. We assume that the estimated model is updated every 10 steps; if these steps are interpreted as trading days this would correspond to every two trading week.
- To make the rolling estimation procedure clear, in each experiment we generate a total dataset of $n + n_2$ values from the true distribution G. The window size is n_2, where the modeller begin by using the data $L_1, \ldots, L_{n_2}$ to form their model F, and make the realized p-values estimates $U_{n_2+i} = F(L_{n_2+i})$, for $i = 1, \ldots, 10$.
- The modeller then roll the dataset forward 10 steps and use the data $L_{11}, \ldots, L_{n_2+10}$ to make realized p-values estimates $U_{n_2+10+i} = F(L_{n_2+10+i})$, for $i = 1, \ldots, 10$; in total the models are thus re-estimated $n/10$ times.

Accuracy of historical simulation

n_2	F | results	Bias	MAE	By10	By25	By33
250	Normal	-2.7	7.0	18.9	0.2	0.0
	t5	-0.1	12.4	28.3	2.8	0.2
	t3	5.0	19.5	31.1	7.8	1.8
	st3	4.5	20.6	33.5	9.7	2.9
500	Normal	-0.9	4.9	6.0	0.0	0.0
	t5	2.7	9.4	13.1	0.1	0.0
	t3	8.6	16.0	16.0	1.1	0.1
	st3	8.7	16.8	17.8	1.7	0.1
1000	Normal	0.1	3.5	1.0	0.0	0.0
	t5	4.1	7.4	3.6	0.0	0.0
	t3	10.5	13.6	5.3	0.1	0.0
	st3	11.1	14.5	5.6	0.0	0.0

Table: Bias and mean absolute error (MAE) (both expressed as percentages) of standard empirical estimator of 97.5% expected shortfall for different sample sizes and different distributions. By10, By25 and by33 give percentage of estimates underestimating expected shortfall by 10%, 25% or 33.3% respectively. Results are based on 10,000 replications.

Multinomial tests historical simulation rejection rate

	n_2	250					500				
G	n | Test	B99	P4	N4	L4	L8	B99	P4	N4	L4	L8
Normal	250	5.6	7.4	6.6	4.3	3.9	3.7	5.2	5.0	6.3	6.2
	500	3.7	6.1	5.8	2.0	1.5	1.6	3.7	3.0	2.9	2.8
	1000	2.7	10.2	9.7	1.2	1.0	0.2	2.3	2.1	1.6	0.2
t5	250	6.2	8.0	7.5	3.9	4.4	3.3	5.3	4.8	4.8	5.2
	500	2.8	6.2	5.5	1.9	1.9	1.8	4.4	4.1	3.6	3.2
	1000	2.4	11.4	10.9	1.4	1.3	0.2	2.6	2.4	2.0	0.8
t3	250	5.7	6.9	6.3	4.0	4.5	2.5	5.8	5.7	5.7	5.6
	500	2.4	5.8	5.1	1.1	1.2	1.7	2.9	2.6	3.5	2.0
	1000	2.6	10.6	10.0	1.6	0.7	0.3	1.9	1.7	1.8	0.8
st3	250	6.1	8.6	8.2	4.3	4.5	2.8	6.0	5.5	6.2	6.2
	500	2.3	6.3	5.4	1.6	1.2	2.2	4.5	3.9	3.9	2.1
	1000	3.5	11.8	11.1	0.8	0.7	0.5	2.1	2.1	1.7	1.2

Table: Rejection rates for Historical Simulation method with various two-sided multinomial tests in the static backtesting experiment. Models are refitted after 10 days. Results are based on 1000 replications.

Realized p-value

- We define the realized p-value, which we denote by U_t, as the probability of observing a realized loss no more extreme than L_t using the forecast model F, i.e. $U_t := F(L_t)$.
- The transformation $F(L_t)$ is also known as the Rosenblatt transformation, where under the null hypothesis, U_t is i.i.d. uniformly distributed in the interval (0,1).
- Assuming that F is continuous, the realized p-value contains sufficient information to test the VaR exception process at all levels, since

$$I_{t,\alpha} = I_{\{L_t > \text{VaR}_\alpha\}} = I_{\{U_t > \alpha\}} .$$

Spectral test for realized p-value

- We borrow ideas from Costanzino & Curran (2015), which proposed to test the Spectral Risk Measures by looking at a weighted integral of VaR exceptions, where they define

$$W_{g,t} := \int_0^1 g(u) I_{\{L_t > \text{VaR}_u\}} \mathrm{d}u,$$

with g being the admissible risk spectrum of the Spectral Risk Measure of interest.

- Since the event $\{L_t > \text{VaR}_\alpha\}$ is same as the event $\{U_t > \alpha\}$, we have that

$$W_{g,t} = \int_0^1 g(u) I_{\{U_t > u\}} \mathrm{d}u.$$

- We can construct a Z-test for $W_{g,t}$, where the Z-test statistic is given by

$$Z_g := \sqrt{\frac{n}{\sigma_g^2}} \left(\hat{\mu}_g - \mu_g \right) ,$$

where $\hat{\mu}_g = \frac{1}{n} \sum_{t=1}^{n} W_{g,t}$, and μ_g and σ_g^2 is the mean and variance of $W_{g,t}$ under H_0.

Spectral test at different weight functions

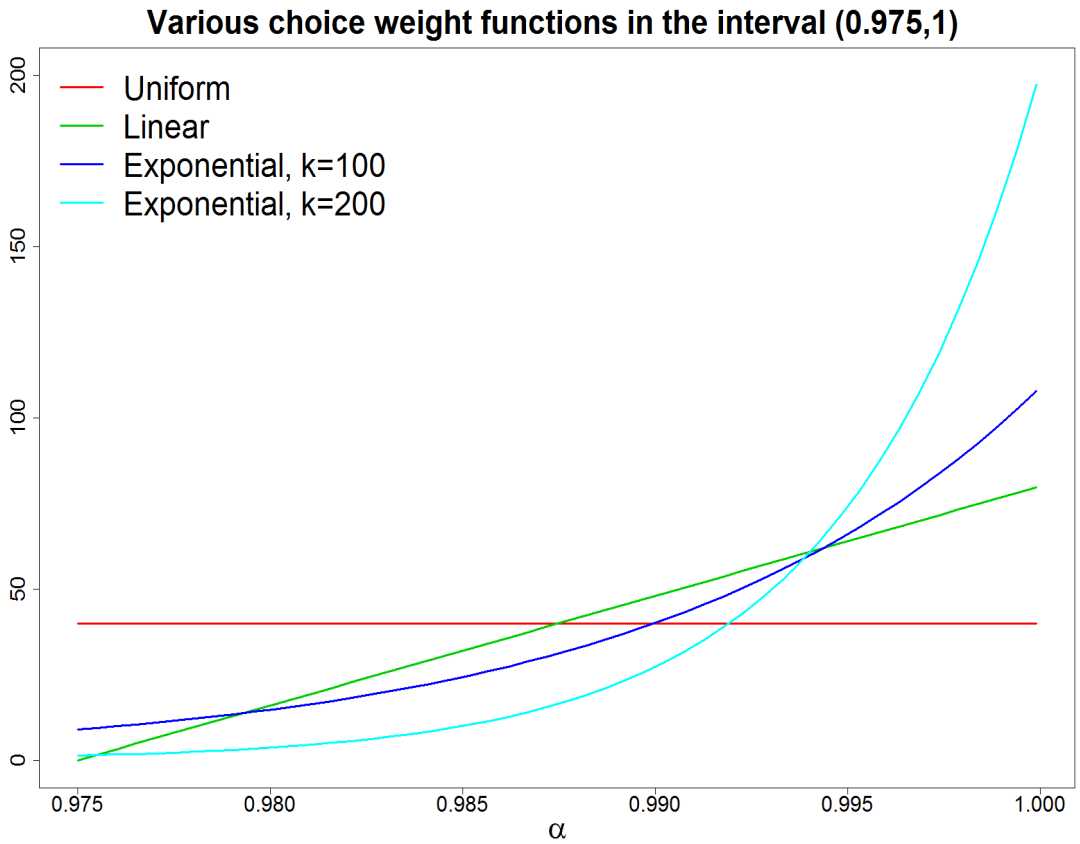

Figure: The uniform, linear, and exponential weight functions, re scaled to unit area, in the interval (0.975,1).

Spectral tests historical simulation rejection rate

	n_2	250				500			
G	n | Test	SP.U	SP.L	SP.E100	SP.E200	SP.U	SP.L	SP.E100	SP.E200
Normal	250	5.1	6.9	7.3	11.4	6.1	6.6	6.5	7.7
	500	2.7	5.1	5.3	9.9	3.8	4.8	4.8	6.3
	1000	1.8	5.3	6.4	16.6	1.0	1.5	1.7	3.6
t5	250	5.4	7.1	8.0	11.2	4.6	6.0	6.5	8.6
	500	2.3	4.4	5.5	9.0	4.2	4.7	4.9	6.5
	1000	2.0	4.7	6.8	15.6	0.7	1.1	1.8	4.2
t3	250	4.9	7.3	8.0	12.1	5.1	5.7	5.9	8.2
	500	2.5	4.9	5.7	12.2	2.8	3.6	4.1	6.0
	1000	2.7	5.9	7.6	16.7	0.8	1.5	1.6	3.9
st3	250	6.8	8.3	8.9	13.6	4.8	4.9	4.9	8.2
	500	3.1	5.3	5.9	11.6	3.7	4.8	4.8	6.3
	1000	2.7	5.5	7.2	17.2	1.1	2.1	2.3	3.7

Table: Rejection rates for Historical Simulation method with various two-sided spectral tests in the static backtesting experiment. Models are refitted after 10 days. Results are based on 1000 replications.

Summary

- We have look at binomial test for testing a single VaR and multinomial test for testing VaR at multiple levels.
- Multinomial tests are more powerful than the binomial test, with LRT having the best performance.
- Multinomial tests are not able to detect underestimation error of historical simulation model.
- Spectral tests with more emphasis on the tail can detect underestimation error of historical simulation model better.

Reference

CHRISTOFFERSEN, P. (1998). Evaluating interval forecasts. International Economic Review **39**.

COSTANZINO, N. & CURRAN, M. (2015). Backtesting general spectral risk measures with application to expected shortfall. Working paper. .

FERNANDEZ, C. & STEEL, M. (1998). On bayesian modelling of fat tails and skewness. Journal of the American Statistical Association **93**, 359âĂŞ371.

NASS, C. (1959). The χ^2-test for small expectations in contingency tables, with special reference to accidents and absenteeism. Biometrika **46**, 365–385.

PEARSON, K. (1932). Experimental discussion of the (χ^2,p) test for goodness of fit. Biometrika **24**, 351âĂŞ381.