Testing Quantile Forecast Optimality

Jack Fosten[1] Daniel Gutknecht[2] **Marc-Oliver Pohle**[3]

[1]King's College London

[2]Goethe University Frankfurt

[3]Heidelberg Institute for Theoretical Studies

12th ECB Conference on Forecasting Techniques
Forecasting "at Risk"

Frankfurt am Main, 12 and 13 June 2023

Motivation

- Quantile forecasts are useful to characterize other distributional features than the central tendency, in particular **risk or uncertainty**.
- They are becoming more and more popular, examples from economics and finance: Value-at-Risk (VaR) and Growth-at-Risk (GaR).
- Usually, such forecasts are issued over **multiple horizons**.
- Often, **multiple quantiles** are of interest, examples: prediction intervals or all deciles to characterize the full distribution.
- Usually, we are interested in evaluating a forecasting approach over all horizons and quantiles of interest. This calls for **multi-horizon, multi-quantile evaluation**.

LITERATURE

- We are interested in **multi-horizon, multi-quantile absolute evaluation of quantile forecasts**.
- Huge literature on absolute evaluation of quantile forecasts (single-horizon, single-quantile), e.g.:
  Christoffersen (1998); Engle and Manganelli (2004); Gaglianone et al. (2011); Nolde and Ziegel (2017)
- Relative multi-horizon evaluation for mean forecasts, extension to quantiles straightforward: Quaedvlieg (2021)
- Absolute multi-horizon evaluation of mean forecasts: Patton and Timmermann (2012)
- **Idea of Paper**: Provide (non-conservative and interpretable) **optimality tests for multi-horizon & multi-quantile forecasts**

CONTRIBUTIONS

- **Tests for autocalibration** based on **quantile Mincer-Zarnowitz (MZ) regressions**:
  - Allow for multi-horizon and multi-quantile forecasts
  - Are based on a finite set of moment (in)equalities
  - Use bootstrap critical values
- **Extensions**:
  - Stronger form of calibration via augmented MZ regressions
  - Multivariate version
- **Simulations** to analyse finite sample performance
- **Application** of the tests in macro and finance

Roadmap of Talk

Optimality and Tests

Extensions

Empirical Application I

Empirical Application II

Conclusion

Set-Up

- Vector time series $\{\mathbf{Z}_t\}_{t=1}^{T}$ containing the target variable y_t and other predictors
- Forecaster's goal: predict the τ-quantile of y_t using information from h periods ago:

$$q_{t,h}\left(\tau|\mathcal{F}_{t-h}\right) = F^{-1}_{y_t|\mathcal{F}_{t-h}}(\tau)$$

  with the information set $\mathcal{F}_{t-h} = \sigma(\{\mathbf{Z}_s; s \leq (t-h)\})$
- Denote an h-period-ahead forecast of $q_{t,h}\left(\tau|\mathcal{F}_{t-h}\right)$ for time t by

$$\widehat{y}_{\tau,t,h}$$

- Forecaster issues forecasts for multiple horizons $h \in \mathcal{H} = \{1, \ldots, H\}$ and multiple quantile ranks $\tau \in \mathcal{T} = \{\tau_1, \ldots, \tau_K\}$.

Quantile Forecast Optimality

- Forecast $\widehat{y}_{\tau,t,h}$ is **optimal w.r.t. the information set** $\mathcal{F}_{t-h}$ if:

$$\widehat{y}_{\tau,t,h} = q_{t,h}\left(\tau|\mathcal{F}_{t-h}\right).$$

- Forecast $\widehat{y}_{\tau,t,h}$ is **autocalibrated** if:

$$\widehat{y}_{\tau,t,h} = q_{t,h}\left(\tau|\sigma\left(\widehat{y}_{\tau,t,h}\right)\right),$$

which is a weaker notion of optimality (Tsyplakov, 2013; Gneiting and Ranjan, 2013)

Autocalibration Test

- Base test on **Mincer-Zarnowitz (MZ)** regressions:

$$y_t = \alpha_h^\dagger(\tau_k) + \beta_h^\dagger(\tau_k)\widehat{y}_{\tau_k,t,h} + \varepsilon_{t,h}(\tau_k)$$

- **Null hypothesis:**

$$H_0^{\text{MZ}} : \{\alpha_h(\tau_k) = 0\} \cap \{\beta_h(\tau_k) = 1\} \text{ for all } h \in \mathcal{H} \text{ and } \tau_k \in \mathcal{T}$$

- Rejecting the null implies systematic errors in the forecasts.
- Our test is interpretable:
  - It shows which quantiles and horizons contribute most strongly to rejection
  - MZ regression lines inform us how forecasts could be improved
- Extends Gaglianone et al. (2011) to multi-horizon and multi-quantile setting

TEST STATISTIC

- Use moment equality framework of Andrews and Soares (2010)
- We observe an evaluation sample of size P, i.e. a scalar-valued time series of observations starting at some point in time $R+1 \in \mathbb{Z}$, $\{y_t\}_{t=R+1}^{T}$, , $(T = P + R)$ and a matrix-valued time series of forecasts, $\left\{ \left(\widehat{y}_{\tau,t,h} \right)_{\tau=\tau_1,\ldots,\tau_K, h=1,\ldots,H} \right\}_{t=R+1}^{T}$
- For each τ_k and h:
  - Estimate the coefficients $\alpha_h(\tau_k)$ and $\beta_h(\tau_k)$ by quantile regression.
  - Define empirical moment $\widehat{m}_s$ either as $\widehat{\alpha}_h(\tau_k)$ or as $(\widehat{\beta}_h(\tau_k) - 1)$
- **Test statistic:**

$$\widehat{U}_{\text{MZ}} = \sum_{s=1}^{\kappa} \left(\sqrt{P} \widehat{m}_s \right)^2,$$

  where P denotes the size of the evaluation sample, $\kappa = 2HK$
- Asymptotic distribution depends on the variance-covariance matrix of the various quantile regression coefficients

Bootstrap Critical Values

- We use moving block bootstrap critical values (CVs) (Künsch, 1989; Gregory et al., 2018), requires choice of block length l
- Resample directly from $\{y_t, \widehat{y}_{\tau_k,t,h}\}_{t=R+1}^{T}$ for each h and τ_k to generate B bootstrap samples

$$\{y_t^b, \widehat{y}_{\tau_k,t,h}^b\}_{t=R+1}^{T}, \quad b = 1, ..., B$$

- Obtain B bootstrap test statistics:

$$\widehat{U}_{\text{MZ}}^b = \sum_{s=1}^{\kappa} \left(\sqrt{P}(\widehat{m}_s^b - \widehat{m}_s)\right)^2$$

  and take critical values as the $1-\alpha$ quantile
- Remarks:
  - Establish asymptotic validity of these CVs
  - Operate under $P/R \to 0$ as $P, R \to \infty$, MC suggests good performance also under equal split $P = R$

Roadmap of Talk

Augmented Mincer Zarnowitz

- Can test a stronger form of optimality relative to a larger information set $\mathcal{I}_{t-h} \subset \mathcal{F}_{t-h}$ than $\sigma(\widehat{y}_{\tau_k,t,h})$
- Propose the **augmented MZ test** using additional regressors $\mathbf{Z}_{t-h}$ from $\mathcal{F}_{t-h}$:

$$y_t = \alpha_h^\dagger(\tau_k) + \widehat{y}_{\tau_k,t,h}\beta_h^\dagger(\tau_k) + \mathbf{Z}_{t-h}'\boldsymbol{\gamma}_h^\dagger(\tau_k) + \varepsilon_{t,h}(\tau_k)$$

and test the composite null hypothesis for all $h \in \mathcal{H}$ and $\tau_k \in \mathcal{T}$:

$$H_0^{\text{AMZ}} : \{\alpha_h^\dagger(\tau_k) = 0\} \cap \{\beta_h^\dagger(\tau_k) = 1\} \cap \{\boldsymbol{\gamma}_h^\dagger(\tau_k) = \mathbf{0}\}$$

- Reject null $\Rightarrow \mathbf{Z}_{t-h}$ contains information which could have improved forecasts

Multiple Target Variables

- Set-up can easily be extended to several target variables $i = 1, ..., G$
- Examples: multiple macro series of interest, VaR of multiple firms in the S&P500 etc.
- We simply extend the MZ regression to multiple time series:

$$y_{i,t+h} = \alpha_{h,i}(\tau_k) + \widehat{y}_{i,\tau_k,t,h}\beta_{h,i}(\tau_k) + \varepsilon_{i,t,h}(\tau_k), \quad i = 1, \ldots, G.$$

and test the composite null hypothesis:

$$H_0^{\text{MMZ}} : \{\alpha_{h,i}(\tau_k) = 0\} \cap \{\beta_{h,i}(\tau_k) = 1\}$$

for all $h \in \mathcal{H}$, $\tau_k \in \mathcal{T}$, $i = 1, \ldots, G$

- Previously $\kappa = 2HK$ moment equalities, now $\kappa = 2GHK$

Roadmap of Talk

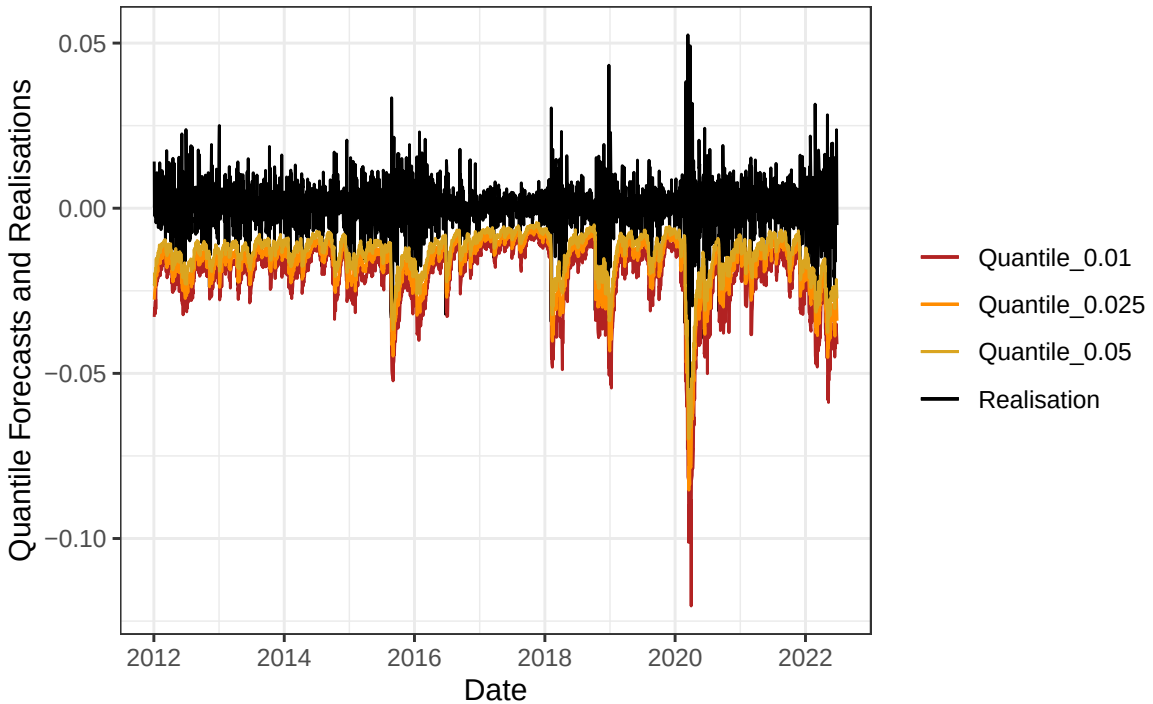
Finance Application

- Backtesting VaR (τ-quantile of return distribution) is a central task in financial risk management
- Typically τ set to be a single level like 1%, 2.5% or 5%
- However, no consensus on this, and measures like expected shortfall (ES) depend on a range of quantiles
- In addition, multi-horizon aspect to risk management: often one day-ahead or cumulative 10-day returns used
- Our test is naturally suited to this set-up

Data and Set-up

- **Target variable:** daily S&P500 returns
- **Model**: GARCH(1,1) of Bollerslev (1986) with GARCH bootstrap of Pascual et al. (2006) to generate multi-step quantile predictions
- **Data source**: Oxford Man Realised Library
- **Data span**: 3rd Jan 2000 to 27th June 2022
- **Estimation scheme**: recursive window
- **Sample sizes**: $T = 5634$ daily observations, initial estimation sample of size $R = 3000$
- **Horizons**: $H = 10$, so $h = 1, ..., 10$
- **Quantile levels**: $\mathcal{T} = \{0.01, 0.025, 0.05\}$
- **Bootstrap:** block length $l = 10$, $B = 1000$ draws
- **Robustness checks**: GJR-GARCH model (Glosten et al., 1993), without Covid-19 period, different bootstrap block lengths l

Results



Stat	90%	95%	99%	p-value
9834.131	5009.153	6569.754	9821.539	0.01

Results - Individual Contributions

Useful to look at individual contributions to this statistic from single quantiles and single horizons

	$\tau = 0.01$	$\tau = 0.025$	$\tau = 0.05$	all
$h = 1$	427.463	81.455	39.217	548.135
$h = 2$	439.467	195.770	50.126	685.363
$h = 3$	672.670	266.256	127.524	1066.450
$h = 4$	591.907	265.840	99.559	957.306
$h = 5$	549.574	431.091	141.886	1122.551
$h = 6$	553.680	431.926	114.260	1099.866
$h = 7$	149.554	291.555	230.722	671.831
$h = 8$	258.922	298.486	223.656	781.063
$h = 9$	560.313	405.563	402.132	1368.008
$h = 10$	497.402	562.498	473.658	1533.558
all	4700.952	3230.439	1902.740	9834.131

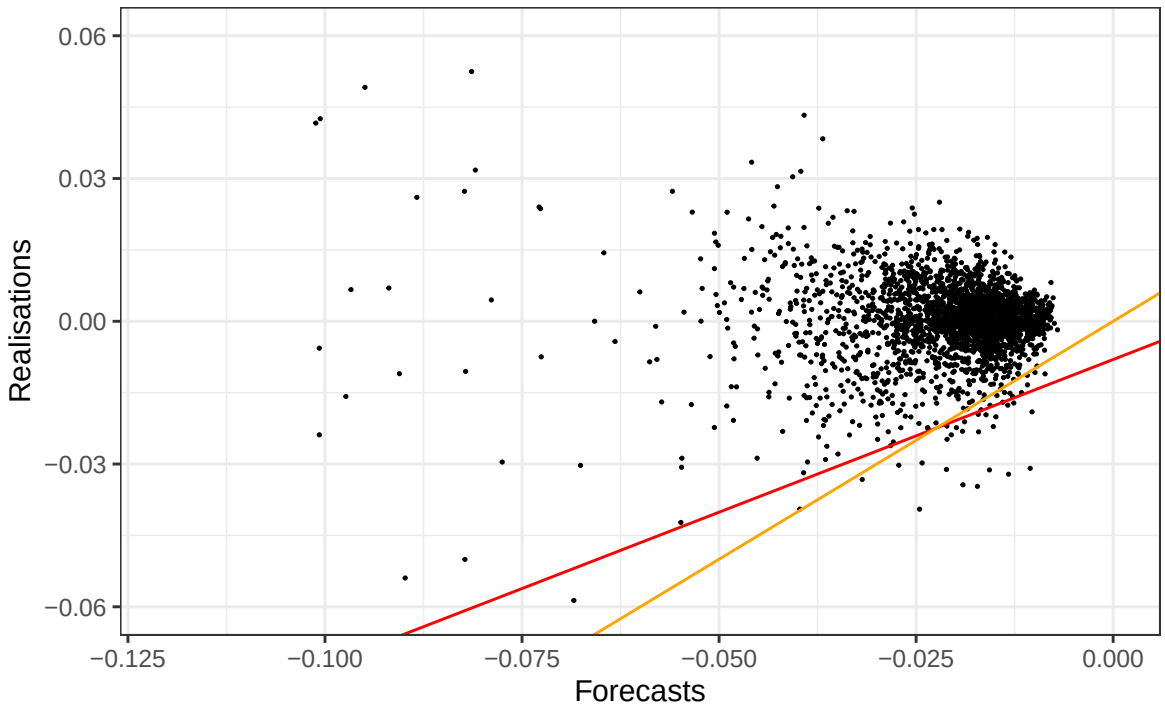
Results - Individual P-Values

What to do without our test? P-values from single quantile and horizon tests:

	$\tau = 0.01$	$\tau = 0.025$	$\tau = 0.05$	all
$h = 1$	0.000	0.092	0.294	0.006
$h = 2$	0.000	0.011	0.161	0.002
$h = 3$	0.008	0.011	0.131	0.001
$h = 4$	0.002	0.023	0.176	0.008
$h = 5$	0.005	0.001	0.141	0.010
$h = 6$	0.010	0.032	0.236	0.012
$h = 7$	0.312	0.073	0.069	0.137
$h = 8$	0.228	0.065	0.091	0.125
$h = 9$	0.030	0.113	0.029	0.028
$h = 10$	0.122	0.044	0.021	0.010
all	0.013	0.011	0.063	0.010

Results - MZ Regression Lines

Compare estimated MZ regression line for $h = 1$ and $\tau = 0.01$ (red) vs. the diagonal (orange), qualitatively same picture emerges for all quantiles and horizons!



Roadmap of Talk

Macro Application

- Quantile forecasting in macro is increasing in popularity since Manzan (2015)
- GaR literature has typically focused on quarterly real GDP growth using NFCI (Adrian et al., 2019)
- More recently applied to quarterly employment, inflation (Adams et al., 2021)
- Explore optimality of model-based forecasts of different U.S. macro variables
- We use monthly variables

Data and Set-up

- **Target variables:** $G = 4$ different targets as in Manzan (2015)
  - Consumer Price Index for All Urban Consumers (CPIAUCSL)
  - Industrial Production: Total Index (INDPRO)
  - All Employees, Total Nonfarm (PAYEMS)
  - Personal Consumption Expenditures Excluding Food and Energy (Chain-Type Price Index) (PCEPILFE)
- **Predictor variables**: Autoregressive term, Chicago Fed National Financial Conditions Index (NFCI)
- **Model**: Linear quantile regression (QADL)
- **Data source**: Federal Reserve Economic Data (FRED)
- **Data span**: 1984M1 to 2019M12
- **Estimation scheme**: recursive window
- **Sample sizes**: $T = 432$ monthly obs split into $R = P = 216$
- **Horizons**: $H = 12$ so $h = 1, ..., 12$
- **Quantile levels**: $\mathcal{T} = \{0.1, 0.25, 0.5\}$
- **Bootstrap:** block length $l = 4$, $B = 1000$ draws

Mincer-Zarnowitz Test Results

	Stat	90%	95%	99%	p-value
Joint	38264.280	28908.454	45259.085	86531.304	0.067
CPIAUCSL	18269.966	18033.852	32452.813	66594.353	0.099
INDPRO	4258.078	7578.204	11224.918	24413.160	0.222
PAYEMS	871.704	1574.085	2060.305	4994.712	0.308
PCEPILFE	14864.532	2316.907	2792.387	3678.394	0.000

- Evidence of miscalibration for inflation series (PCEPILFE and CPIAUCSL), not for real series
- For PCEPILFE and CPIAUCSL the largest contribution to the test statistic comes from quantile level $\tau_k = 0.1$

Augmented Mincer-Zarnowitz Test Results

Re-run Augmented MZ test with additional regressors ($G - 1 = 3$ variables other than target)

	Stat	90%	95%	99%	p-value
CPIAUCSL	21984.030	19794.203	29896.138	57657.304	0.085
INDPRO	5194.690	8722.551	12596.841	27604.813	0.224
PAYEMS	723.354	1494.399	2011.985	4470.360	0.350
PCEPILFE	15648.207	2455.174	2938.071	3801.048	0.000

- No extra regressor seems to improve forecasts for real variables

Roadmap of Talk

Concluding Remarks

- We propose Mincer-Zarnowitz tests for quantile forecast optimality at multiple horizons and multiple quantile levels.
- Test that is straightforward to implement can be extended to:
  1. Augmented Mincer-Zarnowitz test
  2. Multiple time series
- Simulation evidence (not presented) shows tests work well
- Two empirical applications showcase the MZ test and extensions
- Future work: distributional or probabilistic forecasts have become active research field, and test could be adapted using a many moment equality framework (Chernozhukov et al., 2021).

References I

Adams, P. A., T. Adrian, N. Boyarchenko, and D. Giannone (2021). Forecasting macroeconomic risks. *International Journal of Forecasting 37*(3), 1173–1191.

Adrian, T., N. Boyarchenko, and D. Giannone (2019). Vulnerable growth. *American Economic Review 109*(4), 1263–1289.

Andrews, D. and G. Soares (2010). Inference for parameters defined by moment inequalities using generalized moment selection. *Econometrica 78*, 119–157.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics 31*(3), 307–327.

Chernozhukov, V., D. Chetverikov, and K. Kato (2021). Inference on causal and structural parameters using many moment inequalities. *The Review of Economic Studies 86*(5), 1867–1900.

Christoffersen, P. F. (1998). Evaluating interval forecasts. *International economic review*, 841–862.

Engle, R. F. and S. Manganelli (2004). CAViaR: Conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics 22*(4), 367–381.

Gaglianone, W. P., L. R. Lima, O. Linton, and D. R. Smith (2011). Evaluating Value-at-Risk Models via Quantile Regression. *Journal of Business & Economic Statistics 29*(1), 150–160.

Glosten, L. R., R. Jagannathan, and D. E. Runkle (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *The Journal of Finance 48*(5), 1779–1801.

Gneiting, T. and R. Ranjan (2013). Combining predictive distributions. *Electronic Journal of Statistics 7*, 1747–1782.

Gregory, K., S. Lahiri, and D. Nordman (2018). A smooth block bootstrap for quantile regression with time series. *Annals of Statistics 46*(3), 1138–1166.

Künsch, H. (1989). The jackknife and the bootstrap for general stationary observations. *Annals of Statistics 17*(3), 1217–1241.

References II

Manzan, S. (2015). Forecasting the distribution of economic variables in a data-rich environment. *Journal of Business & Economic Statistics 33*(1), 144–164.

Nolde, N. and J. F. Ziegel (2017). Elicitability and backtesting: Perspectives for banking regulation. *The Annals of Applied Statistics 11*(4), 1833–1874.

Pascual, L., J. Romo, and E. Ruiz (2006). Bootstrap prediction for returns and volatilities in GARCH models. *Computational Statistics & Data Analysis 50*(9), 2293–2312.

Patton, A. and A. Timmermann (2012). Forecast rationality tests based on multi-horizon bounds. *Journal of Business and Economic Statistics 30*(1).

Quaedvlieg, R. (2021). Multi-Horizon Forecast Comparison. *Journal of Business & Economic Statistics 39*(1), 40–53.

Tsyplakov, A. (2013). Evaluation of probabilistic forecasts: proper scoring rules and moments. *Available at SSRN 2236605*.