# Do *Any* Economists Have Superior Forecasting Skills?

Ritong Qu    Allan Timmermann    Yinchu Zhu

UC San Diego & Brandeis

Indiana University
01/26/2021

# Super Forecasters

- Media and popular press focus on spectacularly successful forecasts
  - “Black Wednesday” (September 16, 1992): George Soros “broke” the British Pound, earning $1bn
  - US subprime mortgage market crisis (John Paulson)
  - Elephant predicting world cup games
  - Kansas City Quarterback Patrick Mahones predicted to win superbowl in high school yearbook
- Academic research has argued for the existence of “super forecasters” with extraordinary judgment and innate ability to produce accurate forecasts
  - Super forecasters are selected as the best performers from a much larger set
  - skill or luck?

# Testing for Superior skills

- We develop new methods for conducting inference about the existence of forecasters with superior predictive skills in a panel data setting with
  - multiple variables
  - many forecasters
  - many time periods
- Existence of a cross-sectional and time-series dimension for a large set of individual forecasters introduces a high-dimensional multiple hypothesis testing problem
  - many performance statistics are compared
  - Important to control the family wise error rate

# Types of forecasting skills

- We develop new economic hypotheses and tests to identify the nature of the skills that forecasters may possess
  - **Specialist skills**: compare forecasting performance across individual variables or clusters of similar variables
  - **Generalist skills**: compare individual forecasters' average performance across many different variables
  - **Event-specific skills**: compare individual forecaster's average performance across multiple variables in a single period
    - Superior predictive ability during the Global Financial Crisis?

# Specialist vs Generalist skills

- Distinction between specialist and generalist skills is important for understanding sources of forecasting skills
  - Private information unlikely to be available for a large set of macro variables
  - Generalist skills indicative of forecasters' ability to process public information (superior modeling skills)
  - Endogenous information acquisition: Forecasters can choose to focus predominantly on variable-specific information (specialists) or, conversely, on general information (generalists) based on the marginal cost and benefit of information acquisition and processing
  - Example: Mackwiak et al (AER 2009): firms with limited attention rationally pay most attention to the more volatile firm-specific shocks and disregard aggregate shocks

# Methodology

- We develop new Sup tests which apply and extend the bootstrap methods proposed by Chernozhukov et al. (2018)
  - Test the null hypothesis that the benchmark forecast is at least as accurate as an arbitrarily large set of alternative forecasts
  - Our tests can identify superior forecasting skills for *any* economic forecaster for *any* variable or at *any* point in time
    - first tests of equal predictive accuracy conducted over multiple units in a panel setting
  - Bootstrap
    - easy to implement
    - uses studentized test statistic - enhances power of the tests

# Empirical findings

- Bloomberg survey covering monthly forecasts of 14 variables
  - Sample: 1997 - 2019
  - Hundreds of individual forecasters and firms
  - More than 1,000 forecast comparisons in some of our tests
- Empirical findings:
  - Significant evidence that the best forecasters can beat a simple autoregressive benchmark: **Forecasters have skills**
  - Single pairwise forecast comparisons indicate that some individual forecasters can outperform a simple equal-weighted average of their peers
  - Accounting for the multiple hypothesis testing problem, there is little or no significant evidence of superior predictive skills either for individual variables or "on average": **Forecasters do not have any superior skills**

# Comparisons of many forecasts

- Suppose we have $M$ forecasts ($M$ can be large)
- How confident can we be that the best forecast is truly better than some benchmark, given that it is selected from $M$ forecasts?
- **Skill or luck?**
  - Search across multiple forecasts may result in the recovery of a truly good forecast
  - It may also uncover a bad forecast that just happens to be lucky in a given sample
- Tests used in forecast comparisons typically ignore the search that preceded the selection of the top performer

# Single vs. multiple hypothesis testing

- The critical/significance level, $\alpha$, in classical testing controls the type I error, i.e., the probability of discovering a false positive (wrongly rejecting the null)
- In multiple hypothesis testing (MHT), fixing $\alpha$ to test the individual hypotheses will fail to control the overall probability of false positives
  - Suppose $\alpha = 0.05$ and we are testing $m = 20$ hypotheses whose test statistics are independent.
  - The overall Type I error rate is $1 - 0.95^{20} = 0.64$: 64% chance of falsely discovering an anomaly
- Important to account for this issue

# Notations

- Panel of actual and predicted values
  - $i = 1, ...., N$: cross-sectional dimension
  - $t + h = 1, ...., T$ time-series dimension
  - $m = 1, ..., M$ forecasts (forecasters or models)
  - $h \geq 0$: forecast horizon
- $y_{it+h}$: observed value of unit $i$ at time $t + h$
- $\hat{y}_{it+h|t,m}$: forecast of $y_{it+h}$ generated by forecaster (model) $m$ at time $t$
- $e_{i,t+h,m} = y_{i,t+h} - \hat{y}_{i,t+h|t,m}$: forecast error

# Single pairwise comparison of forecast accuracy

- *Loss differential* of forecast $m$, relative to benchmark $m_0$:

$$\Delta L_{i,t+h,m} = L(y_{i,t+h}, \hat{y}_{i,t+h|t,m_0}) - L(y_{i,t+h}, \hat{y}_{i,t+h|t,m})$$

- Under squared error loss

$$\Delta L_{i,t+h,m} = e^2_{it+h,m_0} - e^2_{it+h,m}$$

- Diebold-Mariano (1995) null for a **single** pairwise forecast comparison:

$$H_0^{DM} : E[\Delta L_{i,t+h,m}] = 0$$

- $H_0^{DM}$ can be tested by conducting a robust $t$-test on the time-series sample mean of $\Delta L_{i,t+h,m}$

# Comparing Multiple Forecasts of a Single Variable

- Reality Check null of White (2000):

$$H_0^{RC} : \max_{m \in \{1,\dots,M\}} E[\Delta L_{i,t+h,m}] \leq 0.$$

- RC null tests whether at least one forecast, $m$, is better than the benchmark for a specific variable ($i$)
- RC null is relevant if there is only a single outcome variable ($N = 1$)
  - *ex-ante* we may be interested in studying forecasting performance for a specific unit such as United States in a large cross-country analysis

## Comparing Multiple Forecasts for Individual Forecasters

- Suppose we want to examine the performance of a single forecaster, $m$, relative to the benchmark, $m_0$, across multiple variables ($i = 1, ..., N$) and testing whether this particular forecaster, $m$, is better than the benchmark for *any* of the variables:

$$H_0^m : \max_{i \in \{1,...,N\}} E[\Delta L_{i,t+h,m}] \leq 0.$$

- Under the null, forecaster $m$ does not improve on the benchmark, $m_0$, for *any* of the variables $i = 1, ..., N$
- This null focuses on a single forecaster ($m$) and searches across the set of variables $i = 1, ..., N$
  - dimension of the joint hypothesis test is $N$

# Generalist Skills

- Comparing average performance across multiple variables, we can test for **generalist skills**:

$$H_0^G : \max_{m \in \{1,\ldots,M\}} E[\frac{1}{N}\sum_{i=1}^{N} \Delta L_{i,t+h,m}] \leq 0$$

- Does any forecaster have skills "on average"?
- $H_0^G$ allows individual forecasts, $m = 1, \ldots, M$, to outperform the benchmark for some variables, $i$, as long as the average forecasting performance is worse than for the benchmark, $m_0$

# Specialist Skills

- If a subset of variables with common features can be identified ex-ante, alternatively we can test for domain-specific, specialist skills by comparing the average predictive accuracy for units within this subset (cluster) $C_k$ comprising $N_k < N$ of the variables
- Test for predictive skills for this subset of variables for any of the $M$ forecasters by means of the specialist skill hypothesis

$$H_0^S : \max_{m=1,\ldots,M} E[\frac{1}{N_k} \sum_{i \in C_k} \Delta L_{i,t+h,m}] \leq 0$$

- if $C_k$ only contains a single element, this reduces to the RC null

# Comparing Performance Across Multiple Variables and Multiple Forecasts

- Does there exist *any* variable, $i$, for which *any* of the forecasts, $m$, beats the benchmark?
- Testing this broad "no superior skill" hypothesis requires that we model the distribution of the test statistic obtained by maximizing both over $i$ and $m$:

$$H_0^{NS}: \max_{i\in\{1,\ldots,N\}} \max_{m\in\{1,..,M\}} E[\Delta L_{i,t+h,m}] \leq 0.$$

# Test statistics

- Test statistic for the maximum value of the average loss differential, computed across the $i = 1, ...., N$ cross-sectional units:

$$R_T = \max_{1 \leq i \leq N} \frac{T^{-1/2} \sum_{t+h=1}^{T} I_{i,t+h} \Delta L_{i,t+h}}{\hat{a}_i}$$

- $I_{i,t+h} = \mathbf{1}\{\Delta L_{i,t+h} \text{ is observed}\}$
- $\hat{a}_i > 0$: normalizing scalar
- $\Delta L_{i,t+h} \equiv L_{i,t+h,m_0} - L_{i,t+h,m}$ (drop $m, m_0$)

# Normalizations (choice of $\hat{a}_i$)

We can consider a variety of normalizations:

- No normalization: $\hat{a}_i = 1$ for $1 \leq i \leq n$.
  - No attempt to balance differences in $Var(T^{-1/2} \sum_{t=1}^{T} \Delta L_{i,t+h})$ across $i$
  - $R_T$ is dominated by the largest values of $Var(T^{-1/2} \sum_{t=1}^{T} \Delta L_{i,t+h})$
- Full normalization: $\hat{a}_i = \sqrt{K^{-1} \sum_{j=1}^{K} \left( B_T^{-1/2} \sum_{t \in H_j} (\Delta L_{i,t+h} - \hat{\mu}_i) \right)^2}$
  - corrects the cross-sectional differences in scale of $T^{-1/2} \sum_{t=1}^{T} \Delta L_{i,t+h}$
- Partial normalization: $\hat{a}_i = \sqrt{T^{-1} \sum_{t+h=1}^{T} (\Delta L_{i,t+h} - \hat{\mu}_i)^2}$ with $\hat{\mu}_i = T^{-1} \sum_{t=1}^{T} \Delta L_{i,t+h}$
  - corrects for different scales in the unconditional variance of $Var(\Delta L_{i,t+h})$

# Multiplier Bootstrap

- Critical values for $R_T$ can be based on a multiplier bootstrap procedure
- $\xi_{t+h}$ : set of i.i.d $N(0,1)$ variables used to construct the statistic

$$R_T^* = \max_{1 \le i \le N} R_{i,T}^*,$$

where

$$R_{i,T}^* = \frac{T^{-1/2} \sum_{t+h=1}^{T} \xi_{t+h} I_{i,t+h} \Delta L_{i,t+h}}{\hat{a}_i}$$

- Theoretical justification uses Theorem B.1 in Chernozhukov et al. (2018)
- $W_{k,t+h} = \Delta L_{k,t+h} - E(\Delta L_{k,t+h})$.

## Assumption 1

*Suppose that the following conditions hold:*
*(1) The distribution of* $W_{t+h}$ *does not depend on* $t$.
*(2)*$P(\max_{1\leq t+h\leq T}\|W_{t+h}\|_{\infty}\leq D_T)=1$ *for some* $D_T\geq 1$.
*(3)* $\{W_{t+h}\}_{t+h=1}^{T}$ *is* $\beta$*-mixing with mixing coefficient* $\beta_{\text{mixing}}(\cdot)$.
*(4)* $c_1\leq E\left(k^{-1/2}\sum_{t+h=s+1}^{s+k}W_{j,t+h}\right)^2,\ E\left(k^{-1/2}\sum_{t+h=s+1}^{s+k}W_{j,t+h}\right)^2\leq C_1$ *for any* $j$, $s$ *and* $k$.
*(5)* $T^{1/2+b}D_T\log^{5/2}(\mathcal{N}T)\lesssim B_T\lesssim T^{1-b}/(\log\mathcal{N})^2$ *and* $\beta_{\text{mixing}}(s)\lesssim\exp(-b_1 s^{b_2})$ *for some constant* $b,b_1,b_2>0$.
*(6) There exist a nonrandom vector* $a=(a_1,...,a_{\mathcal{N}})'\in\mathbb{R}^{\mathcal{N}}$ *and constants* $\kappa_1,\kappa_2>0$ *such that* $\kappa_1\leq a_j\leq\kappa_2$ *for all* $1\leq j\leq\mathcal{N}$ *and* $\max_{1\leq j\leq\mathcal{N}}|\hat{a}_j-a_j|=o_P(1/\log\mathcal{N})$.

# Assumption 1

- Part (1): strict stationarity - can be relaxed at expense of more technicalities in proof
- Part (2): bound on the tail behavior of loss differences. Needed for the high-dimensional bootstrap and Gaussian approximation
- Part (3): $\beta$-mixing (routine assumption)
- Part (4): Loss differences of all variables should be roughly of the same order
- Part (5): Rate conditions. We allow $N >> T$

# Distribution of test statistic

**Theorem 1**

*Suppose Assumption 1 holds. Under*

$$H_0: \max_{1\leq i\leq N} \max_{1\leq m\leq M} E\left[L(y_{i,t+h}, \hat{y}_{i,t+h|t,m_0}) - L(y_{i,t+h}, \hat{y}_{i,t+h|t,m})\right] \leq 0,$$

*we have*

$$\limsup_{T\to\infty} P\left(\tilde{R}_T > \tilde{Q}^*_{T,1-\alpha}\right) \leq \alpha,$$

*where* $\tilde{Q}^*_{T,1-\alpha}$ *is the* $(1-\alpha)$ *quantile of* $\tilde{R}^*_T$ *conditional on the data.*

- Theorem 1 implies that the probability of a false discovery is at most $\alpha$

# Corollary 1

- Let $A = \{i : \mu_i > 0\}$, so $A$ is the set of units, $i$, for which an alternative forecast, $m$, beats the benchmark, $m_0$
- $\hat{A}$: Estimated set of superior forecasters:

$$\hat{A} = \left\{ i : \frac{T^{-1/2} \sum_{t=1}^{T} \Delta L_{i,t+h}}{\hat{a}_i} > Q^*_{T,1-\alpha} \right\}.$$

- With probability at least $1 - \alpha$, $\hat{A}$ only selects variables for which the alternative forecast outperforms the benchmark:

## Corollary 1

*Suppose Assumption 1 holds. Then, for $A$ and $\hat{A}$ defined above,*

$$\limsup_{T \to \infty} P\left(\hat{A} \subseteq A\right) \geq 1 - \alpha.$$

# Interpretation

- Theorem 1 implies that the probability of a false discovery is asymptotically at most $\alpha$
- With probability at least $1 - \alpha$, $\hat{A}$ only selects variables for which the alternative forecast outperforms the benchmark.

# Comparing performance in a single period

- If $N$ is large, we can exploit the cross-sectional dimension to test the "event skill" null that no forecaster has a better cross-sectional average performance than the benchmark in a *single* period, $t+h$:

$$H_0^{ES}: \max_{m \in \{1,\ldots,M\}} E[\frac{1}{N}\sum_{i=1}^{N} \Delta L_{i,t+h,m}] \leq 0.$$

- or in *any* time period:

$$H_0^{ES'}: \max_{t+h \in \{1,\ldots,T\}} \max_{m \in \{1,\ldots,M\}} E[\frac{1}{N}\sum_{i=1}^{N} \Delta L_{i,t+h,m}] \leq 0.$$

- Tests use the average cross-sectional loss differentials

$$\hat{\mu}_{t+h,m} = N^{-1}\sum_{i=1}^{N} \Delta L_{i,t+h,m}$$

# Testing for “event-specific skills”

- Need to model cross-sectional dependencies in loss differences
- Let $f_{t+h}$ be latent factors and assume a factor structure for the forecast errors:

$$e_{i,t+h,m} = \lambda'_{i,m} f_{t+h} + u_{i,t+h,m}$$

- Rule out strong cross-sectional dependencies
  - Idiosyncratic terms assumed to be independent conditional on the factor structure

# Assumption 2

## Assumption 2

*Let $\mathcal{F}$ be the $\sigma$-algebra generated by $\{f_{t+h}\}_{1\leq t+h\leq T}$ and $\{\lambda_{i,m}\}_{1\leq i\leq n,\ 0\leq m\leq M}$. Assume that conditional on $\mathcal{F}$, $\{u_i\}_{i=1}^{n}$ is independent across $i$ and $E(u_i \mid \mathcal{F}) = 0$, where $u_i = \{u_{i,t+h,m}\}_{1\leq t+h\leq T, 1\leq m\leq M} \in \mathbb{R}^{T\times M}$.*

# Test statistic

- Test statistic:

$$Z = \max_{(t+h,m)\in\hat{A}} \frac{\sqrt{N}\overline{\Delta L}_{t+h,m}}{\sqrt{N^{-1}\sum_{i=1}^{N}\widetilde{\Delta L}_{i,t+h,m}^{2}}}.$$

- Critical values for this test statistic can be obtained from a bootstrap

$$Z_* = \max_{(t+h,m)\in\hat{A}} \frac{N^{-1/2}\sum_{i=1}^{N}\varepsilon_i\widetilde{\Delta L}_{i,t+h,m}}{\sqrt{N^{-1}\sum_{i=1}^{N}\widetilde{\Delta L}_{i,t+h,m}^{2}}},$$

with multipliers $\varepsilon_i \sim N(0,1)$ generated independently of the data

# Theorem 2

**Theorem 2**

*Under the assumed factor structure (Assumption 2) and the conditional null*

$$H_0 : \max_{(t+h,m)\in\mathcal{A}} E\left(\frac{1}{N}\sum_{i=1}^{N}\Delta L_{i,t+h,m} \mid \mathcal{F}\right) \leq 0,$$

*we have*

$$\limsup_{N\to\infty} P\left(Z > Q^{*}_{N,1-\alpha,Z}\right) \leq \alpha,$$

*where $Q^{*}_{n,1-\alpha,Z}$ is the $(1-\alpha)$ quantile of $Z_{*}$ conditional on the data*

# Monte Carlo simulations: Size

- Forecast errors obey a factor structure
- **non-studentized** test statistic
  - tends to be undersized for large $N$ and $T$, particularly when $M$ is also large
- **studentized** test statistic
  - good size for small-to-modest values of $N$ and $M$, but tends to be undersized for large $N, T, M$
  - undersizing is strongest for $\alpha = 0.05$
  - test statistic is over-sized for small $N, T$
- Power can go from 10-15% for the non-studentized to 70-80% for the studentized test statistic

# Table A1, size

$\alpha = 0.1$

Without studentization | With studentization

$M = 2$

| $N \setminus T$ | 25 | 50 | 100 | 200 | 25 | 50 | 100 | 200 |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.117 | 0.135 | 0.111 | 0.113 | 0.117 | 0.131 | 0.116 | 0.117 |
| 10 | 0.109 | 0.112 | 0.105 | 0.108 | 0.126 | 0.113 | 0.077 | 0.081 |
| 25 | 0.115 | 0.112 | 0.093 | 0.112 | 0.141 | 0.098 | 0.076 | 0.073 |
| 50 | 0.086 | 0.117 | 0.097 | 0.100 | 0.122 | 0.087 | 0.054 | 0.059 |
| 100 | 0.087 | 0.099 | 0.109 | 0.086 | 0.148 | 0.074 | 0.058 | 0.045 |

$M = 10$

| $N \setminus T$ | 25 | 50 | 100 | 200 | 25 | 50 | 100 | 200 |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.121 | 0.158 | 0.143 | 0.119 | 0.113 | 0.134 | 0.113 | 0.088 |
| 10 | 0.134 | 0.143 | 0.121 | 0.104 | 0.155 | 0.101 | 0.075 | 0.068 |
| 25 | 0.127 | 0.155 | 0.123 | 0.130 | 0.170 | 0.083 | 0.043 | 0.049 |
| 50 | 0.105 | 0.135 | 0.123 | 0.095 | 0.196 | 0.072 | 0.044 | 0.033 |
| 100 | 0.104 | 0.100 | 0.077 | 0.070 | 0.231 | 0.079 | 0.030 | 0.031 |

# Table A2, size adjusted critical values

$\alpha = 0.1$

| | Without studentization | | | | With studentization | | | |
|---|---|---|---|---|---|---|---|---|
| | $M = 2$ | | | | $M = 2$ | | | |
| $N \setminus T$ | 25 | 50 | 100 | 200 | 25 | 50 | 100 | 200 |
| 1 | 0.084 | 0.076 | 0.084 | 0.088 | 0.076 | 0.080 | 0.080 | 0.088 |
| 10 | 0.096 | 0.088 | 0.096 | 0.092 | 0.076 | 0.092 | 0.124 | 0.120 |
| 25 | 0.096 | 0.092 | 0.108 | 0.092 | 0.068 | 0.108 | 0.132 | 0.128 |
| 50 | 0.112 | 0.092 | 0.104 | 0.104 | 0.088 | 0.112 | 0.140 | 0.152 |
| 100 | 0.112 | 0.104 | 0.096 | 0.112 | 0.064 | 0.124 | 0.136 | 0.168 |
| | $M = 10$ | | | | $M = 10$ | | | |
| $N \setminus T$ | 25 | 50 | 100 | 200 | 25 | 50 | 100 | 200 |
| 1 | 0.084 | 0.072 | 0.076 | 0.084 | 0.092 | 0.080 | 0.092 | 0.112 |
| 10 | 0.076 | 0.080 | 0.088 | 0.096 | 0.060 | 0.100 | 0.120 | 0.128 |
| 25 | 0.088 | 0.072 | 0.088 | 0.088 | 0.060 | 0.112 | 0.148 | 0.156 |
| 50 | 0.096 | 0.084 | 0.088 | 0.104 | 0.052 | 0.116 | 0.156 | 0.180 |
| 100 | 0.100 | 0.104 | 0.116 | 0.124 | 0.040 | 0.116 | 0.172 | 0.188 |

# Monte Carlo simulations: Power

- Randomly select 20% of the competing forecasts and add $\left(2T^{-1}\log(MN)\right)^{1/8}$ to their forecast errors, which then have larger MSE than the baseline forecasts
    - size-adjusted critical values used to study power
- General conclusion:
    - Studentized test statistic has far better power than the non-studentized test
    - Power can go from 10-15% for the non-studentized to 70-80% for the studentized test statistic

# Table A3, power

$\alpha = 0.1$

Without studentization | With studentization

$M = 2$

| $N \setminus T$ | 25 | 50 | 100 | 200 | 25 | 50 | 100 | 200 |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.095 | 0.087 | 0.105 | 0.090 | 0.082 | 0.083 | 0.095 | 0.093 |
| 10 | 0.218 | 0.240 | 0.305 | 0.304 | 0.462 | 0.515 | 0.682 | 0.775 |
| 25 | 0.216 | 0.208 | 0.280 | 0.255 | 0.609 | 0.712 | 0.892 | 0.950 |
| 50 | 0.218 | 0.196 | 0.225 | 0.234 | 0.770 | 0.834 | 0.955 | 0.994 |
| 100 | 0.197 | 0.205 | 0.180 | 0.244 | 0.776 | 0.890 | 0.976 | 0.999 |

$M = 10$

| $N \setminus T$ | 25 | 50 | 100 | 200 | 25 | 50 | 100 | 200 |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.559 | 0.468 | 0.551 | 0.706 | 0.415 | 0.393 | 0.500 | 0.687 |
| 10 | 0.182 | 0.207 | 0.213 | 0.274 | 0.733 | 0.827 | 0.945 | 0.998 |
| 25 | 0.184 | 0.204 | 0.212 | 0.258 | 0.818 | 0.905 | 0.991 | 1.000 |
| 50 | 0.213 | 0.241 | 0.218 | 0.279 | 0.829 | 0.904 | 0.990 | 1.000 |
| 100 | 0.241 | 0.261 | 0.281 | 0.365 | 0.815 | 0.925 | 0.999 | 1.000 |

# Bloomberg Data

- Bloomberg conducts monthly surveys of financial and macroeconomic variables
- We focus on forecasts of 14 variables
  - "Release date": date when the official data source publishes the actual value of a variable
  - "Observation date" (earlier): end of the period covered by the survey

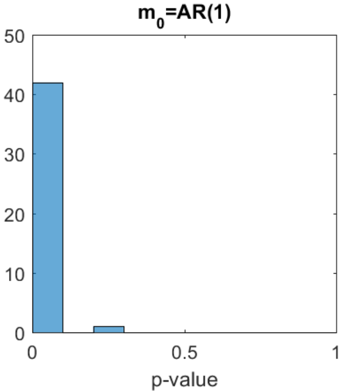
# Summary statistics

Table: Summary of Bloomberg survey variables

| Viariable name | Description | Frequency | Time series observation | Number of forecasters | Number of firms | Number of firms>5 forecasts |
|---|---|---|---|---|---|---|
| AHE | Average hourly earnings | monthly | 111 | 104 | 86 | 38 |
| CPI | CPI | monthly | 197 | 178 | 134 | 67 |
| ETSL | Existing homes sales | monthly | 171 | 215 | 162 | 92 |
| FDTR | Fed Funds rate | 8 times/year | 169 | 544 | 395 | 88 |
| GDP | GDP | monthly | 254 | 309 | 221 | 134 |
| GDPC | GDP Personal Consumption | monthly | 193 | 167 | 130 | 50 |
| IP | Industrial Production | monthly | 252 | 288 | 204 | 121 |
| NFP | Nonfarm payrolls | monthly | 254 | 324 | 234 | 153 |
| NHS | New home sales | monthly | 251 | 273 | 196 | 103 |
| NHSPA | Building permits | monthly | 202 | 205 | 150 | 69 |
| NHSPS | Housing starts | monthly | 252 | 278 | 198 | 99 |
| PCEC | PCE Core | monthly | 180 | 164 | 121 | 45 |
| PCE | PCE | monthly | 181 | 154 | 118 | 65 |
| UN | Unemployment | monthly | 253 | 308 | 224 | 149 |

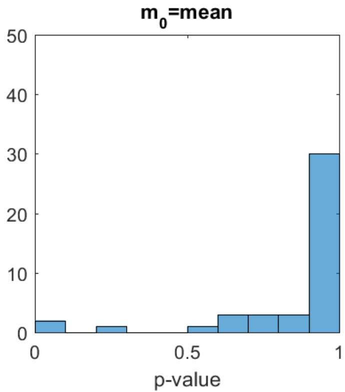
# Sup tests for individuals covering multiple variables

(a) Firms vs AR(1)



# Sup tests for individuals covering multiple variables (cont.)

(a) Firms vs mean

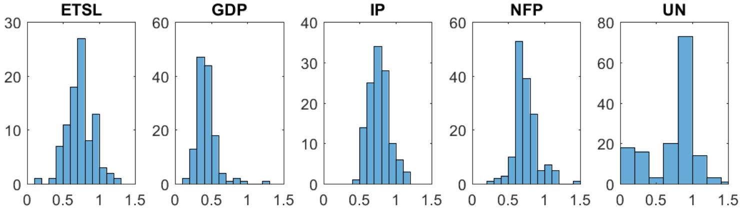


# Bloomberg Data: Diebold-Mariano tests

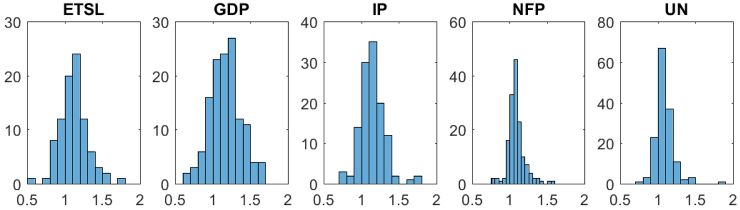
- Pair-wise Diebold Mariano tests
  - A majority of forecasters are more accurate than the forecasts from the AR(1) model for most variables–often significantly so
    - Few individual forecasters are significantly *more* accurate than the equal-weighted (EW) mean
    - Many individual forecasters are significantly *worse* than the EW mean

# RMSE Ratios

(a) Firms vs AR(1)




(b) Firms vs mean



# Diebold-Mariano tests

Table: Distribution of DM test tstatistics. Firm level forecasters vs. AR(1) or mean.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Panel A: firm forecasts vs AR(1)** | | | | | | | | |
| | AHE | CPI | ETSL | FDTR | GDP | IP | NFP | UN |
| tstat<-1.645 | 0 | 0 | 0 | 0 | 2 | 0 | 1 | 4 |
| -1.645<tstat<0 | 2 | 0 | 6 | 3 | 0 | 9 | 13 | 15 |
| 0<tstat<1.645 | 9 | 17 | 48 | 61 | 14 | 32 | 42 | 78 |
| tstat>1.645 | 27 | 50 | 38 | 24 | 118 | 80 | 97 | 52 |
| **Panel B: firm forecasts vs mean** | | | | | | | | |
| tstat<-1.645 | 14 | 28 | 30 | 9 | 56 | 45 | 58 | 63 |
| -1.645<tstat<0 | 17 | 28 | 39 | 41 | 51 | 57 | 72 | 59 |
| 0<tstat<1.645 | 6 | 11 | 19 | 34 | 24 | 18 | 19 | 26 |
| tstat>1.645 | 1 | 0 | 4 | 4 | 3 | 1 | 4 | 1 |
| total | 38 | 67 | 92 | 88 | 134 | 121 | 153 | 149 |

# Tests of Reality Check null

Individual forecasters vs. AR(1):

- $m_0 = AR(1)$, $m = forecasters$: Strongly reject $H_0^{RC}$
  - number of significantly better forecasters is much smaller than suggested by the pair-wise DM tests
- $m_0 = forecasters$, $m = AR(1)$: Fail to find a single rejection of $H_0^{RC}$

Individual forecasters vs. mean

- $m_0 = mean$, $m = forecasters$: Across 14 variables, only two cases (one, each, for GDPC and NFP) in which $H_0^{RC}$ is rejected
- $m_0 = forecasters$, $m = mean$: many more rejections of $H_0^{RC}$, particularly for GDP, IP and UN

# Sup tests for individual variables

Table: Sup tests for predictive dominance

| | AHE | CPI | ETSL | FDTR | GDP | IP | NFP | UN | Average |
|---|---|---|---|---|---|---|---|---|---|
| **Panel A: $m_0$ = AR(1), $m_1$ = firm forecasts** | | | | | | | | | |
| pval | 0.00 | 0.00 | 0.02 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| n rejections | 18 | 18 | 4 | 2 | 51 | 15 | 27 | 9 | 49 |
| **Panel B: $m_0$ = firm forecasts, $m_1$ = AR(1)** | | | | | | | | | |
| pval | 1.00 | 1.00 | 1.00 | 0.99 | 0.94 | 1.00 | 0.79 | 0.53 | 1.00 |
| n rejections | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Panel C: $m_0$ = mean, $m_1$ = firm forecasts** | | | | | | | | | |
| | 0.94 | 0.98 | 0.20 | 0.53 | 1.00 | 0.97 | 0.01 | 0.84 | 1.00 |
| n rejections | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| **Panel D: $m_0$ = firm forecasts, $m_1$ = mean** | | | | | | | | | |
| pval | 0.03 | 0.00 | 0.02 | 0.53 | 0.00 | 0.01 | 0.03 | 0.00 | 0.00 |
| n rejections | 2 | 5 | 4 | 0 | 7 | 6 | 5 | 9 | 36 |
| n forecasters | 38 | 67 | 92 | 88 | 134 | 121 | 153 | 149 | 121 |

# Sup tests across subsets of variables

| **Panel A:** $m_0$ **= AR(1),** $m_1$ **= firm forecasts** | Inflation | Housing market | Growth | Labor | Funds rate |
|---|---|---|---|---|---|
| p-value | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| no. rejections | 33 | 17 | 66 | 36 | 7 |
| **Panel B:** $m_0$ **= firm forecasts,** $m_1$ **= AR(1)** | | | | | |
| p-value | 0.95 | 1.00 | 0.99 | 0.97 | 0.99 |
| no. rejections | 0 | 0 | 0 | 0 | 0 |
| **Panel C:** $m_0$ **= mean,** $m_1$ **= firm forecasts** | | | | | |
| p-value | 0.98 | 1.00 | 0.72 | 0.04 | 0.52 |
| no. rejections | 0 | 0 | 0 | 1 | 0 |
| **Panel D:** $m_0$ **= firm forecasts,** $m_1$ **= mean** | | | | | |
| p-value | 0.00 | 0.02 | 0.00 | 0.00 | 0.53 |
| no. rejections | 12 | 1 | 16 | 17 | 0 |
| no. forecasters | 87 | 123 | 147 | 155 | 88 |

# Testing for superior skills for *any* forecasters

- 1,001 pairwise comparisons
- $m_0 = AR(1)$, $m = forecasters$: We identify 49 individual forecasters who are significantly more accurate than the AR(1) model for at least one variable ($p - val = 0.00$)
- $m_0 = forecasters$, $m = AR(1)$: Fail to reject the reverse null – that all forecasters are at least as accurate for all variables as the AR(1) forecasts ($p - val = 0.65$)
- $m_0 = mean$, $m = forecasters$: Only a single instance where an individual forecaster beats the EW average ($p - val = 0.03$)
- $m_0 = forecasters$, $m = mean$: six cases where the EW average is significantly better than individual forecasters

# Sup tests across variables and forecasters (I)

Table: Sup tests for equal predictive accuracy. Multiple variable, multiple forecasts

| | Benchmark vs. firm forecasters | | Benchmark vs. individual forecasters | |
|---|---|---|---|---|
| **Panel A: partial studentization** | | | **Panel B: partial studentization** | |
| | $m_0$ =AR(1) | Reverse | $m_0$ =AR(1) | Reverse |
| pval | 0.00 | 0.65 | 0.00 | 0.71 |
| n rejections | 49 | 0 | 47 | 0 |
| **Panel C: no studentization** | | | **Panel D: no studentization** | |
| | $m_0$ =AR(1) | Reverse | $m_0$ =AR(1) | Reverse |
| pval | 0.27 | 0.99 | 0.19 | 0.99 |
| n rejections | 0 | 0 | 0 | 0 |
| **Panel E: moment selection** | | | **Panel F: moment selection** | |
| | $m_0$ =AR(1) | Reverse | $m_0$ =AR(1) | Reverse |
| pval | 0.04 | 0.73 | 0.04 | 0.80 |
| n rejections | 1 | 0 | 1 | 0 |

# Sup tests across variables and forecasters (II)

Table: Sup tests for equal predictive accuracy. Multiple variable, multiple forecasts

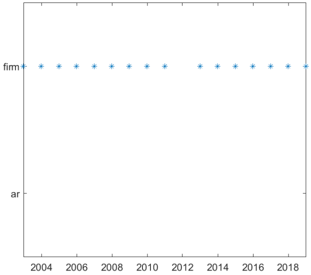
| | Benchmark vs. firm forecasters | | Benchmark vs. individual forecasters | |
|---|---|---|---|---|
| | **Panel G: partial studentization** | | **Panel H: partial studentization** | |
| | $m_0$ =mean | Reverse | $m_0$ =mean | Reverse |
| pval | 0.03 | 0.01 | 0.02 | 0.00 |
| n rejections | 1 | 6 | 1 | 7 |
| | **Panel I: no studentization** | | **Panel J: no studentization** | |
| | $m_0$ =mean | Reverse | $m_0$ =mean | Reverse |
| pval | 0.91 | 0.15 | 0.94 | 0.15 |
| n rejections | 0 | 0 | 0 | 0 |
| | **Panel K: moment selection** | | **Panel L: moment selection** | |
| | $m_0$ =mean | Reverse | $m_0$ =mean | Reverse |
| pval | 0.44 | 0.03 | 0.49 | 0.03 |
| n rejections | 0 | 1 | 0 | 1 |

# Bloomberg Data: Event skills

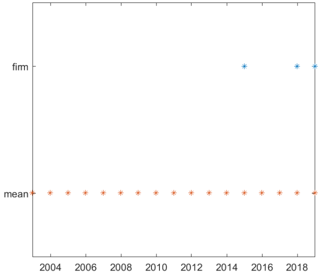
- Compute cross-sectional average test statistics using non-overlapping 12-month blocks
  - 16 of 17 years where at least one individual forecaster is significantly more accurate than the AR(1) benchmark
  - zero years where the reverse holds and at least one forecaster is less accurate than the AR(1) benchmark
  - 3 years where at least one forecaster is more accurate than the EW average
  - EW average is more accurate than at least one individual forecaster every single year

# Sup test for individual years

(a) Firms vs AR(1)

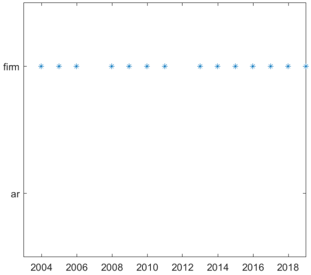



(b) Firms vs mean

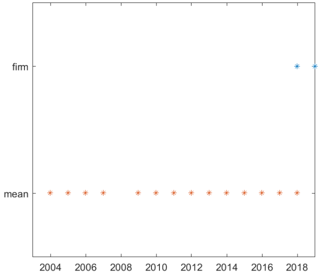


# Sup test across all years

(a) Firms vs AR(1)




(b) Firms vs mean



# Term Structure of Uncertainty

- International Monetary Fund's World Economic Outlook (WEO) forecasts of real GDP growth and inflation across the world's economies
- WEO is published twice each year: April (Spring, or $S$) and October (Fall, or $F$) for the current-year ($h = 0$) and next-year ($h = 1$) periods:
  - $\{h = 1, S; h = 1, F; h = 0, S; h = 0, F\}$.
- Compare WEO forecasts at long versus short horizons
- WEO forecasts only involve pair-wise comparisons ($M = 1$)
  - Cross-sectional dimension (country-level) is large: $N = 180$ countries
  - Time-series dimension: 1990-2016 ($T = 27$ years)

# WEO forecasts across horizons

- Ordering the WEO forecasts from longest to shortest horizon,

$$E[e^2_{h=0,F}] \leq E[e^2_{h=0,S}] \leq E[e^2_{h=1,F}] \leq E[e^2_{h=1,S}].$$

- Define the loss differential for forecasts generated at short and long horizons, $t - h_S$ and $t - h_L$ for $h_L > h_S$:

$$\Delta L_{i,t,h_L \to h_S} = (y_{i,t} - \hat{y}_{i,t|t-h_S})^2 - (y_{i,t} - \hat{y}_{i,t|t-h_L})^2.$$

- Test the null that, for each country, $i$, the forecast is at least as accurate at the short horizon, $h_S$, as it is at the long horizon, $h_L > h_S$:

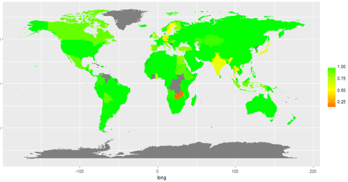
$$H_0 : \max_{i \in \{1,\ldots,N\}} (E[\Delta L_{i,t,h_L \to h_S}]) \leq 0.$$

- Test the reverse null:

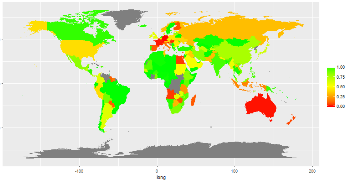
$$H_0 : \max_{i \in \{1,\ldots,N\}} (E[\Delta L_{i,t,h_L \to h_S}]) \leq 0.$$

# Sup test across different horizons, inflation

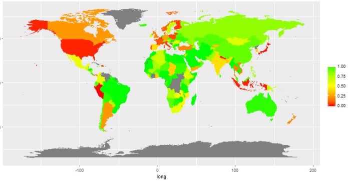
(a) h=1, S vs. h=1, F



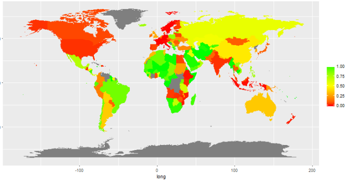

(b) h=1, F vs. h=0, S




(c) h=0, S vs. h=0, F




(d) h=1, S vs. h=0, F



# Table 6, Sup tests across different horizons

| | | | | |
|---|---|---|---|---|
| Panel A: GDP, $m_0$ = short horizon, $m_1$ = long horizon | | | | |
| h=1, S vs. h=1, F | h=1, F vs. h=0, S | h=0, S vs. h=0, F | h=1, S vs. h=0, F | |
| 0.992 | 0.999 | 0.925 | 1 | |
| Panel B: GDP, $m_0$ = long horizon, $m_1$ = short horizon | | | | |
| h=1, S vs. h=1, F | h=1, F vs. h=0, S | h=0, S vs. h=0, F | h=1, S vs. h=0, F | |
| 0.091 | 0.002 | 0.008 | 0.005 | |
| Brazil | Switzerland | Chile | Argentina | Lebanon |
| Italy | Venezuela | Israel | Brazil | Panama |
| Portugal | | Italy | Comoros | Peru |
| | | Japan | Congo, DRC | Portugal |
| | | Spain | Guyana | Switzerland |
| | | St. Kitts Nevis | Haiti | Tunisia |
| | | Switzerland | Israel | United States |
| | | Ukraine | Italy | Venezuela |
| | | United Kingdom | Kenya | Zimbabwe |

# Table 6 (cont.): Sup tests across different horizons

| **Panel C: Inflation, $m_0$ = short horizon, $m_1$ = long horizon** | | | | |
|---|---|---|---|---|
| h=1, S vs. h=1, F | h=1, F vs. h=0, S | h=0, S vs. h=0, F | h=1, S vs. h=0, F | |
| 0.316 | 0.944 | 1.000 | 0.998 | |
| **Panel D: Inflation, $m_0$ = long horizon, $m_1$ = short horizon** | | | | |
| h=1, S vs. h=1, F | h=1, F vs. h=0, S | h=0, S vs. h=0, F | h=1, S vs. h=0, F | |
| 0.127 | 0.000 | 0.000 | 0.000 | |
| | Angola | Belgium | Angola | Italy |
| | Australia | Dominican Republic | Austria | Kenya |
| | Cyprus | Finland | Bangladesh | Lithuania |
| | Egypt | France | Belarus | Luxembourg |
| | Finland | Indonesia | Belgium | Malaysia |
| | France | Italy | Canada | Mongolia |
| | Germany | Japan | Cyprus | Mozambique |
| | Hungary | Lithuania | Denmark | New Zealand |
| | Luxembourg | Nepal | Dominican Republic | Norway |
| | Madagascar | Peru | Egypt | Portugal |
| | New Zealand | Poland | Estonia | Romania |
| | Slovak Republic | Portugal | Ethiopia | Spain |
| | Slovenia | Singapore | Finland | Sweden |
| | Spain | United States | France | Switzerland |
| | Switzerland | | Germany | Thailand |
| | Zimbabwe | | Ghana | United States |
| | | | Guatemala | Zambia |
| | | | India | Zimbabwe |
| | | | Indonesia | |

# Table 7: Sup tests for subsets of variables, long vs. short horizon, GDP

**Panel A: GDP Growth**

| | world | ae | emde | lics | eeur | dasia | lac | menap | cis | ssa |
|---|---|---|---|---|---|---|---|---|---|---|
| h=1,S vs. h=1,F | 0.01 | 0.01 | 0.01 | 0.07 | 0.14 | 0.42 | 0.00 | 0.14 | 0.04 | 0.05 |
| no. rejections | 3 | 3 | 1 | 1 | 0 | 0 | 2 | 0 | 1 | 1 |
| h=1,F vs. h=0,S | 0.00 | 0.01 | 0.00 | 0.07 | 0.20 | 0.03 | 0.00 | 0.12 | 0.03 | 0.05 |
| no. rejections | 2 | 6 | 1 | 1 | 0 | 1 | 3 | 0 | 4 | 2 |
| h=0,S vs. h=0,F | 0.01 | 0.00 | 0.04 | 0.10 | 0.02 | 0.03 | 0.01 | 0.04 | 0.00 | 0.07 |
| no. rejections | 9 | 14 | 3 | 0 | 1 | 2 | 5 | 3 | 2 | 1 |
| h=1,S vs. h=0,F | 0.00 | 0.00 | 0.00 | 0.00 | 0.03 | 0.03 | 0.00 | 0.01 | 0.01 | 0.00 |
| no. rejections | 20 | 15 | 15 | 9 | 2 | 4 | 12 | 5 | 3 | 10 |
| no. countries | 186 | 36 | 150 | 58 | 12 | 28 | 32 | 23 | 12 | 43 |

# Table 7: Sup tests for subsets of variables, long vs. short horizon, inflation

**Panel B: Inflation**

| | world | ae | emde | lics | eeur | dasia | lac | menap | cis | ssa |
|---|---|---|---|---|---|---|---|---|---|---|
| h=1,S vs. h=1,F | 0.13 | 0.16 | 0.11 | 0.06 | 0.15 | 0.13 | 0.43 | 0.33 | 0.43 | 0.04 |
| no. rejections | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 2 |
| h=1,F vs. h=0,S | 0.00 | 0.00 | 0.03 | 0.03 | 0.00 | 0.05 | 0.03 | 0.01 | 0.04 | 0.01 |
| no. rejections | 17 | 13 | 7 | 3 | 6 | 1 | 3 | 4 | 2 | 4 |
| h=0,S vs. h=0,F | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.06 | 0.02 | 0.14 |
| no. rejections | 15 | 16 | 5 | 2 | 3 | 5 | 5 | 1 | 2 | 0 |
| h=1,S vs. h=0,F | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 |
| no. rejections | 38 | 28 | 20 | 9 | 5 | 9 | 7 | 7 | 6 | 8 |
| no. countries | 185 | 36 | 149 | 58 | 12 | 28 | 31 | 23 | 12 | 43 |

# Table 8: Sup tests across different horizons (AEs)

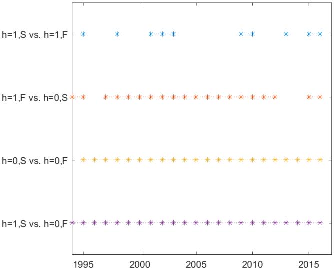
| **Panel A: GDP, $m_0$ = short horizon, $m_1$ = long horizon** | | | |
|---|---|---|---|
| h=1, S vs. h=1, F | h=1, F vs. h=0, S | h=0, S vs. h=0, F | h=1, S vs. h=0, F |
| 0.975 | 0.755 | 1.000 | 1.000 |
| **Panel B: GDP, $m_0$ = long horizon, $m_1$ = short horizon** | | | |
| 0.007 | 0.007 | 0.003 | 0.002 |
| Italy | Canada | Belgium | Belgium |
| Japan | Hong Kong SAR | Canada | Canada |
| Portugal | Luxembourg | Cyprus | Cyprus |
| | Portugal | Estonia | Finland |
| | Switzerland | France | France |
| | United States | Israel | Germany |
| | | Italy | Greece |
| | | Japan | Hong Kong SAR |
| | | Latvia | Ireland |
| | | New Zealand | Israel |
| | | Portugal | Italy |
| | | Spain | Japan |
| | | Switzerland | Luxembourg |
| | | United Kingdom | Malta |
| | | | Portugal |
| | | | Switzerland |
| | | | United States |

# Table 8: Sup tests across different horizons (AEs)

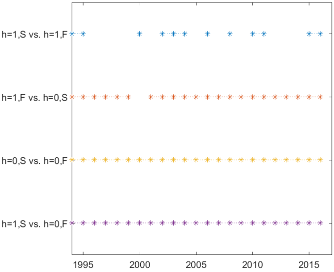
| | | | | |
|---|---|---|---|---|
| **Panel C: Inflation,** $m_0$ = **short horizon**, $m_1$ = **long horizon** | | | | |
| h=1, S vs. h=1, F | h=1, F vs. h=0, S | h=0, S vs. h=0, F | h=1, S vs. h=0, F | |
| 0.888 | 1.000 | 1.000 | 1.000 | |
| **Panel D: Inflation,** $m_0$ = **long horizon**, $m_1$ = **short horizon** | | | | |
| h=1, S vs. h=1, F | h=1, F vs. h=0, S | h=0, S vs. h=0, F | h=1, S vs. h=0, F | |
| 0.151 | 0.000 | 0.001 | 0.000 | |
| | Australia | Belgium | Austria | Netherlands |
| | Cyprus | Canada | Belgium | New Zealand |
| | Finland | Denmark | Canada | Norway |
| | France | Finland | Cyprus | Portugal |
| | Germany | France | Czech Republic | Singapore |
| | Italy | Germany | Denmark | Slovak Republic |
| | Luxembourg | Italy | Estonia | Slovenia |
| | New Zealand | Japan | Finland | Spain |
| | Slovak Republic | Lithuania | France | Sweden |
| | Slovenia | New Zealand | Germany | Switzerland |
| | Spain | Norway | Ireland | United Kingdom |
| | Switzerland | Portugal | Italy | United States |
| | | Singapore | Japan | |
| | | Slovak Republic | Korea | |
| | | United Kingdom | Lithuania | |
| | | United States | Luxembourg | |

# Sup test for individual years (inflation)

(a) GDP

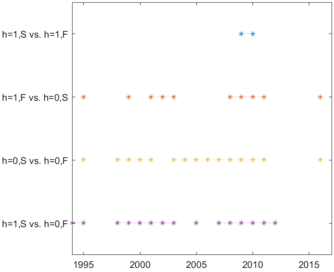



(b) CPI

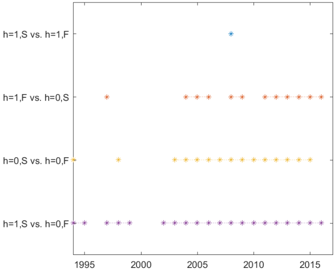


# Sup test across all years (inflation)

(a) GDP




(b) CPI



# Conclusions

- We develop new panel forecast methods for testing if individual forecasts are significantly more accurate–after accounting for the multiple hypothesis testing problem–than a benchmark forecast for at least one
  - outcome variable
  - forecaster (model)
  - time-period
- Tests build on the Chernozhukov (2018) bootstrap approach
  - important to extend this to use studentized test statistics
- We test for specialist, generalist, or event-specific forecasting skills
  - We can identify the forecasters, variables, and time periods for which forecasters possess superior skills
- Empirically, we find that forecasters are skilled (beat a simple, robust time-series model), but do not, on the whole, possess superior skills relative to a simple equal-weighted average