Testing the accuracy of the out-of-sample forecast

(Updated Spring 2021)

Chaoyi Chen
Institute of MNB, Corvinus University of Budapest

Empirical Financial Econometrics

Outlines

- Testing the accuracy of the out-of-sample forecast ⏩ Jump [Online Lecture]
- Tests for Forecast Data Mining ⏩ Jump [Self-study if you have interest :)]

Testing the accuracy of the out-of-sample forecast

- Testing for forecast bias
- Testing for predictability of forecast errors
- Measure of forecast accuracy
- Two common loss functions
- The Diebold-Mariano (DM) test
- Forecasting encompassing tests
- Model averaging

Testing for forecast bias

- Recall that when we use $E_t y_{t+1}$ as our forecast for y_{t+1}, it is conditionally unbiased by construction. i.e. if we set $f_t \equiv E_t y_{t+1}$, then $E_t y_{t+1} = f_t$ by construction.
- However, in practice, it's not so easy to find an unbiased forecast.
- In particularly, we must
  1. choose a model for y_{t+1} that we will use to construct $E_t^m y_{t+1}$, where "m" refers to the model used to calculate $E_t y_{t+1}$. Now,

$$f_t^m = E_t^M y_{t+1} \begin{cases} = E_t y_{t+1}, & \text{if } m \text{ is correct model} \\ \neq E_t y_{t+1}, & m \text{ is wrong model.} \end{cases}$$

$$E_t y_{t+1} = E_t y_{t+1} = E_t^m y_{t+1} + \underbrace{(E_t y_{t+1} - E_t^m y_{t+1})}_{\text{forecast bias}}$$

  2. Estimate parameters of model to use $\widehat{y}_{t+1}^m$ in place of $E_t^m y_{t+1}$.

Testing for forecast bias Cont.

- The upshot is that in practice we don't know if our forecasts are unbiased or not & we might want to test to see if they are.
- Likewise, if we have two competing forecast models & we found that one was biased and the other unbiased, then we discard biasded forecast and keep the unbiased forecast.
- Let $\widehat{y}^m_{t+h|t}$ be our h period forecast using model m.
- We want to test:

$$H_0: \ E_t y_{t+h} = \widehat{y}^m_{t+h|t} \text{ (model m provides unbiased forecast)} \quad (1)$$

$$H_A: \ E_t y_{t+h} \neq \widehat{y}^m_{t+h|t} \text{ (model m provides biased forecast).}$$

- Consider the following regression (known as a Miner-Zarnowitz regression)

$$y_{t+h} = \beta_0 + \beta_1 \widehat{y}^m_{t+h|t} + \varepsilon_{t+h}; \ E_t \varepsilon_{t+h} = 0. \quad (2)$$

- Calculate $E_t y_{t+h}$ in (2): $E_t y_{t+h} = \beta_0 + \beta_1 \widehat{y}^m_{t+h|t}$.
- So, $E_t y_{t+h} = \widehat{y}^m_{t+h|t}$ if $\beta_0 = 0$ and $\beta_1 = 1$.

Testing for forecast bias Cont.

- That means that we can reformulate the hypothesis in (1) as:

$$\left.\begin{aligned} y_{t+h} &= \beta_0 + \beta_1 \widehat{y}^m_{t+h|t} + \varepsilon_{t+h};\ E_t \varepsilon_{t+h} = 0 \\ H_0 :\ (\beta_0, \beta_1) &= (0,1)\ (\widehat{y}^m_{t+h|t} \text{ unbiased}) \\ H_A :\ (\beta_0, \beta_1) &\neq (0,1)\ (\widehat{y}^m_{t+h|t} \text{ biased}) \end{aligned}\right\} \tag{3}$$

- This can be implemented via an F test.

Testing for predictability of forecast errors

- Suppose re-arrange (3) as:

$$\underbrace{\underbrace{(y_{t+h} - \widehat{y}^m_{t+h|t})}_{\widehat{e}^m_{t+h|t}}}_{\text{forecast error}} = \underbrace{\beta_0}_{\gamma_0} + \underbrace{(\beta_1 - 1)}_{\gamma_1} \widehat{y}^m_{t+h|t} + \varepsilon_{t+h}.$$

$$\left.\begin{aligned} \widehat{e}^m_{t+h|t} &= \gamma_0 + \gamma_1 \widehat{y}^m_{t+h|t} + \varepsilon_{t+h}, \\ H_0 &: \ (\gamma_0, \gamma_1) = 0, \\ H_A &: \ (\gamma_0, \gamma_1) \neq 0. \end{aligned}\right\} \tag{4}$$

- The hypothesis in (4) is identical to the one in (3) since

$$\beta_1 = 1 \Longleftrightarrow \gamma_1 \equiv (\beta_1 - 1) = 0.$$

- If forecast is unbiased, then forecast error is unpredictable.

Testing for predictability of forecast errors Cont.

- Nevertheless, (4) suggests a slightly different interpretation.
- If $\gamma_1 \neq 0$, then ut us possible to predict or forecast errors ($\widehat{e}^m_{t+h|t}$).
- But if our forecast $\widehat{y}^m_{t+h|t}$ satisfies $E_t y_{t+h} = \widehat{y}^m_{t+h|t}$, then it should not be possible to predict the forecast error using information available at the time of forecast since:

$$E_t \widehat{e}^m_{t+h} = E_t \left[y_{t+h} - \widehat{y}^m_{t+h|t} \right] = E_t y_{t+h} - \widehat{y}^m_{t+h|t} = 0 \ (\text{if } E_t y_{t+h} = \widehat{y}^m_{t+h|t}).$$

- Intuitively, if we can predict the forecast error, we should be able to improve our forecast so as to reduce forecast error.
- For 1-horizon ($h = 1$) forecast the condition $E_t \widehat{e}^m_{t+1} = 0$ further implies that.
- This principle is more general than the test in (3).

Testing for predictability of forecast errors: a generalized view

- If any variable known at the time of the forecast (time t) can predict the forecast error, then our forecast is not optimal and could be improved. So we may generalize (3) to

$$\left.\begin{aligned} \widehat{e}^{m}_{t+h|t} &= \gamma_0 + \gamma_1 x_t + \varepsilon_{t+h} \\ H_0: \ (\gamma_0, \gamma_1) &= (0,0) \ \text{(unpredictable forecast errors} \\ &\iff \text{optimal forecast)} \\ H_A: \ (\gamma_0, \gamma_1) &\neq (0,0) \ \text{(predictable forecast errors} \\ &\iff \text{sub-optimal forecast)} \end{aligned}\right\} \quad (5)$$

- Here x_t be any variable (or vector of several variables available at time t). This might include
  1. $y_t, y_{t-1}, y_{t-2} \cdots$
  2. $z_t, z_{t-1}, z_{t-2}, \ldots \longleftarrow$ another variable
  3. $\widehat{y}^{m}_{t}$ - our forecast as in (3)
  4. $\widehat{y}^{m*}_{t}$ - a forecast from another model

Testing for predictability of forecast errors: a generalized view Cont.

- In fact, for the one-period ahead forecast ($h = 1$), if the forecast is unbiased ($E_t \widehat{e}^m_{t+1|t} = 0$), then it must be serially uncorrelated ($cov(\widehat{e}^m_{t+1|t}, \widehat{e}^m_{s+1|s}) = 0$) for $s \neq t$.
- That means that we can use tests such as the Box-Pierce & Ljung-Box statistics to test for unpredictability of the forecast errors.
- Likewise, we could estimate an ARMA model for $\widehat{e}^m_{t+1|t}$ & reject unpredictability if any of the coefficients showed up significant.
- But, for longer horizon forecasts ($h > 1$), be careful because

$$E_t \widehat{e}^m_{t+h|t} = 0 \text{ does } \underline{\text{NOT}} \text{ imply, e.g, } cov(\widehat{e}^m_{t+h|t}, \widehat{e}^m_{t+h-1|t}) = 0$$

Measure of forecast accuracy

- What if we have two forecasts neither of which is unbiased and both of whose forecast errors are predictable.
- On one hand, we know there is a better forecast model out there.
- On the other hand, we need to choose one of our two methods for immediate use.
- How do we compare two sub-optimal forecasts?
- Need to define a loss function $\tilde{L}(\widehat{y}^m_{t+h|t}, y_{t+h})$, which tells us the cost of our forecasting error.
- Usually, the loss function can be written in terms of the forecast error

$$\tilde{L}(\widehat{y}^m_{t+h|t}, y_{t+h}) = L\left[(\widehat{y}_{t+h|t} - \widehat{y}^m_{t+h|t})\right] = L(\widehat{e}_{t+h|t}).$$

Measure of forecast accuracy cont.

- It make sense that $L(\widehat{e}^m_{t+h})$ should obtain a minimum value of zero at $\widehat{e} = 0$.
- A symmetric loss function penalizes over and undershooting equally
  ⏩ Graph
- An asymmetric loss function penalizes overshooting more than undershooting or vice versa ⏩ Graph
  e.g. Forecasting tornado: better safe than sorry

	Actual		
Forecast		Tornado	No Tonardo
	Tornado	✓	Bad
	No Tornado	Really Bad	✓

Measure of forecast accuracy cont.

- A loss function can be derived from a decision problem in which the forecast will play a role.
  e.g.: An investor might use a loss function derived from a mean variance trade-off

- This is what should be done when making forecasts for a specific decision.
- On the other hand, when reporting your forecast to other economists, it is more common to use a few commonly accepted loss functions given in the following slides

Loss function definition

- $L(\widehat{e}^m_{t+h|t}) \equiv$ loss using $\widehat{y}^m_{t+h}$ to forecast y_{t+h}.
- $E[L(\widehat{e}^m_{t+h|t})] \equiv$ expected or population loss using model - m to forecast y_{t+h}.

$$\bar{L} = \frac{1}{T^2}\sum_{t=T-T_2+1}^{T} L(\widehat{e}^m_{t+h|t})$$

$\equiv$ Sample average loss using model-m to forecast the T_2 points in our forecast sample $(y_{T-T_2+h}, ..., y_{T+h})$.

Two common loss functions: MSE

- Mean Squared Error (MSE)

$$L_{MSE}(\widehat{e}_{t+h|t}) = \widehat{e}^2_{t+h|t} \text{ (squared prediction error)}$$

$$E\left[L_{MSE}(\widehat{e}_{t+h|t})\right] = E\left[\widehat{e}^2_{t+h|t}\right] \text{ (expected squared prediction error)}$$

$$\bar{L}_{MSE} = \frac{1}{T_2}\sum_{t=T-T_2+1}^{T} \widehat{e}^2_{t+h|t} \text{ (mean-squared prediction error).}$$

Two common loss functions: MAE

- Mean Absolute Error (MAE)

$$L_{MAE}(\widehat{e}_{t+h|t}) = |\widehat{e}_{t+h|t}| \text{ (absolute value of prediction error)}$$

$$EL_{MAE} = E|\widehat{e}_{t+h|t}| \text{ (expected absolute prediction error)}$$

$$\bar{L}_{MAE} = \frac{1}{T_2}\sum_{t=T-T2+1}^{T}|\widehat{e}_{t+h|t}| \text{ (mean absolute prediction error)}$$

Two common loss functions: Remark

- $E[L_{MSE}]$ & $E[L_{MAE}]$ are population value & are ultimate interest.
- $\bar{L}_{MSE}$ & $\bar{L}_{MAE}$ are the sample averages that we observe & report in practice.
- In squaring $\widehat{e}$, the MSE also squares units of observation. e.g. $\$ \longrightarrow \$^2$. To get back tothe original units its common to report $RMSE = \sqrt{\bar{L}_{MSE}}$ instead/ in addition to $MSE = \bar{L}_{MSE}$.
- By T_2, I mean the number of out-of-sample forecast errors you use to calculate the loss. I do not mean the original sample size.
- In practice, it is the empirical loss function $\bar{L}$ that is calculated.
- Usually, the idea is to compare this function forecasts from two or more forecast methods to find which produces the lowest loss.
- Informally, this is often referred to as a "horse race".

Two common loss functions: Remark cont.

- See previous lecture note for instructions on creating the sample of T_2, out-of-sample forecast errors to which the loss functions may apply.
- The MSE squares the units of obervation (e.g. \$ become \$^{2}). Converting to RMSE converts back to the original units.

Testing to see if one forecast out-performs another

- If we compare two forecasts by their MSE, RMSE, or MAE, one forecast will always "win" the horse race.
- Does this mean that the winner is inherently better (according to our loss function)?
- Or did the winner simply have good luck on these particular forecasts?
- In other word, is the difference in, say MSE, significant?
- To do this, we need a test to compare the two forecasts.

The Diebold-Mariano (DM) test

- Diebold-Mariano test refers to: Enders p.84-86 & Dielbold p.299-300.
- Goal: compare two forecasts using a loss function & using a statistical test to identify significant differences.
  Notation I use below:

Name	Model 1	Model 2	Difference
model	m_1	m_2	
forecast horizon	h=1	h=1	
forecast errors	$\widehat{e}^{m1}_{t+1\|t}$	$\widehat{e}^{m2}_{t+1\|t}$	
forecast errors loss	$L(\widehat{e}^{m1}_{t+1\|t})$	$L(\widehat{e}^{m2}_{t+1\|t})$	$d_{t+1} \equiv L(\widehat{e}^{m1}_{t+1\|t}) - L(\widehat{e}^{m2}_{t+1\|t}))$
expected loss	$E[L(\widehat{e}^{m1}_{t+1\|t})]$	$E[L(\widehat{e}^{m2}_{t+1\|t})]$	$E[d_{t+1}] = E[L(\widehat{e}^{m1}_{t+1\|t})] - E[L(\widehat{e}^{m2}_{t+1\|t})]$
sample average loss	$\bar{L}^{m1} = \sum_{t=T_2+1}^{T} L(\widehat{e}^{m1}_{t+1\|t})$	$\bar{L}^{m1} = \frac{1}{T_2}\sum_{t=T-T_2+1}^{T} L(\widehat{e}^{m2}_{t+1\|t})$	$\bar{d} = \frac{1}{T_2}\sum_{t=T-T_2+1}^{T} d_{t+1} = \bar{L}^{m1} - \bar{L}^{m2}$

- New definitions are:
  1. $d_{t+1} = L(\widehat{e}^{m_1}_{t+1|t}) - L(\widehat{e}^{m_2}_{t+1|t})$
  2. $E[d_{t+1}] = E[L(\widehat{e}^{m_1}_{t+1|t})] - E[L(\widehat{e}^{m_2}_{t+1|t})]$
  3. $\bar{d} = \bar{L}^{m_1} - \bar{L}^{m_2} =$ sample mean (d_{t+1})

The Diebold-Mariano (DM) test: hypothesis

- Null hypothesis: EQUALLY GOOD FORECAST MODELS
- Alternative hypothesis: One of the two forecasts is better (according to L)

$$H_0: \; E[L^{m_1}] = E[L^{m_2}] \iff E[d] = 0$$
$$H_A: \; E[L^{m_1}] \neq E[L^{m_2}] \iff E[d] \neq 0$$

- Note: It is the population loss function we want to compare but the sample loss function that we observe

The Diebold-Mariano (DM) test: intuition

- Small positive or negative values of $\bar{d} = \bar{L}^{m_1} - \bar{L}^{m_2}$ support H_0 large absolute values support H_A.
- How large? Need to know distribution of $\bar{d} = \bar{L}^{m_1} - \bar{L}^{m_2}$.
- DM showed that $\bar{d} \underset{approx}{\sim} N(d, var(\bar{d}))$ for large T_2
- So, when $H_0 : d = 0$ holds $\bar{d} \underset{\substack{approx \\ H_0}}{\sim} N(0, var(\bar{d}))$
- which implies:

$$\frac{\bar{d}}{\sqrt{var(\bar{d})}} \underset{\substack{approx \\ H_0}}{\sim} N(0,1) \text{ for large } T_2$$

The Diebold-Mariano (DM) test: intuition cont.

- $var(\bar{d})$ unkown
- Suppose we have a good estimator, say $\widehat{var}(\bar{d})$ for $var(\bar{d})$. Then, $\widehat{var}(\bar{d}) \underset{approx}{=} var(\bar{d})$ for large T_2.
- Then, $z \equiv \frac{\bar{d}}{\widehat{var}(\bar{d})} \underset{\substack{approx \\ H_0}}{\sim} N(0,1)$
- Reject for $|z| > z_{\frac{\alpha}{2}}$ ⏩ Graph
- All that's left is to specify the estimator for the variance, $\widehat{\sigma_{\alpha}^2}$
- This is the most complicated part, so the explanation will be skipped.

The Diebold-Mariano (DM) test: HAC

- DM suggest what is commonly referred to as a heteroskedasticity and autocorrelation (HAC) variance estimator.
- This simplist form of this estimator is known as the truncated estimator & given by

$$T\widehat{var}(\bar{d}) = \sum_{r=-M}^{M} \widehat{\gamma}_{\bar{d}}(r) = \widehat{\gamma}_{\bar{d}}(0) + 2\sum_{h=1}^{M} \widehat{\gamma}_{\bar{d}}(h)$$

$$M = T_2^{1/3}$$

$$\widehat{\gamma}_{\bar{d}}(r) = \text{sample covariance}(d_t, d_{t+r})$$

$$\widehat{var}(\bar{d}) = \frac{1}{T}[\widehat{\gamma}_{\bar{d}}(0) + 2\sum_{r=1}^{M} \widehat{\gamma}_{\bar{d}}(r)]$$

$$z = \frac{\bar{d}}{\sqrt{\frac{1}{T}[\widehat{\gamma}_{\bar{d}}(0) + 2\sum_{r=1}^{M} \widehat{\gamma}_{\bar{d}}(r)]}}$$

Forecasting encompassing tests

- Want to forecast at time t
- Have two forecasts from two models:
  $\widehat{y}^{m_1}_{t+h|t}$ from model 1
  $\widehat{y}^{m_2}_{t+h|t}$ from model 2
- Have generated a sample of out-of-sample forecasts as explained in previous lecture.
- Now consider running a regression of your observed data (y_{t+h}) on both forecasts

$$y_{t+h} = \beta_0 + \beta_1 \widehat{y}^{m_1}_{t+h|t} + \beta_2 \widehat{y}^{m_2}_{t+h|t} + u_{t+h} \tag{6}$$

- Now, consider the hypothesis

$$H_0 : \ \beta_2 = 0 \tag{7}$$

$$H_A : \ \beta_2 \neq 0 \tag{8}$$

- If $\beta_2 = 0$ then $\widehat{y}^{m_2}_{t+h|t}$ has no additional predictive power for y_{t+h} after controlling for $\widehat{y}^{m_1}_{t+h|t}$

Forecasting encompassing tests cont.

- I.e. $\beta_2 = 0$ tells us that the second forecast adds no value relative to the first
- Technically speaking, we would say that the information content in the first forecast encompassing the information content in the second forecast.
  Intuitively ⏩ Graph
- Under $H_A : \beta_2 \neq 0$ the second forecast can be used to improve the first forecast.
- That doesn't necessarily mean that the 2nd forecast is better than the first, just that it has something to add to or improve upon the first forecast.
  Intuitively, ⏩ Graph
- Of course we can also test the hypothesis:

$$\left.\begin{aligned} H_0 : \ \beta_1 = 0 \\ H_A : \ \beta_1 \neq 0 \end{aligned}\right\} \tag{9}$$

Model averaging

- If the null hypothesis in both encompassing tests, (7) & (9), are rejected that implies that both forecasts have something to contribute above & beyond what's already in the other forecast.
- So why not use both?
- In particular, with $\beta_1 \neq 0$ & $\beta_2 \neq 0$, (6) suggests that the linear combination of forecasts given by:

$$\widehat{y}_{t+h|t} = \beta_0 + \beta_1 \widehat{y}^{m_1}_{t+h|t} + \beta_2 \widehat{y}^{m_2}_{t+h|t} \tag{10}$$

  should out-of-perform either individual forecast.
- (10) is an example of model averaging, in which we combine (kind of average) two or more forecasts, to make a new & hopefully more reliable forecast.
- In practice, we don't know β_0, β_1, β_2 so we must instead estimate the regression in (10) & use:

$$\widehat{y}_{t+h|t} = \widehat{\beta}_0 + \widehat{\beta}_1 \widehat{y}^{m_1}_{t+h|t} + \widehat{\beta}_2 \widehat{y}^{m_2}_{t+h|t} \tag{11}$$

Model averaging cont.

- In practice, we estimate $\widehat{\beta}_0$, $\widehat{\beta}_1$, $\widehat{\beta}_2$ using the "psuedo" out-of-sample forecasts. We create following previous lecture notes & then use the coefficients to construct our real out-of-sample forecasts.
- Estimate the β_s with error does add some noise to the forecast
- A simpler form of model averaging which doesn't add noise but also doesn't pick it weights optimally is:

$$\widehat{y}_{t+h|t} = \frac{1}{2}\widehat{y}^{m_1}_{t+h|t} + \frac{1}{2}\widehat{y}^{m_2}_{t+h|t} \tag{12}$$

- This is quite literally a model average and if explains why the terminology model averaging is employed

Model averaging cont.

- Note that both the encompassing tests & model averaging techniques all extend naturally to 3 or more forecasts
- Finally, note that for $h > 1$, μ_t in (6) is likely serially correlated & robust (i.e. HAC) standard errors should be employed for the encompassing tests

$$
\begin{aligned}
E[\varepsilon_t^2] &= \alpha_0 + \alpha_1 E[\varepsilon_{t-1}^2] \\
&= \alpha_0 + \alpha_1 2E[\varepsilon_t^2] \\
\implies & (1 - \alpha_1 L) E[\varepsilon_t^2] = \alpha_0 \\
\implies & E[\varepsilon_t^2] = \frac{1}{1 - \alpha_1 L} \alpha_0 \\
&= \frac{\alpha_0}{1 - \alpha_1 L}
\end{aligned}
$$

Tests for Forecast Data Mining

- Comparing multiple forecasts to a benchmark
- Data mining
- The basic multiple test problem
- How to pick the appropriate critical value?
- Procedures to select

Comparing multiple forecasts to a benchmark

- References for the further reading include:
  1. White, H. (2000). A reality check for data snooping. Econometrica, 68(5), 1097-1126.
  2. Romano, J. P., amp; Wolf, M. (2005). Stepwise multiple testing as Formalized data snooping. Econometrica, 73(4), 1237-1282.
  3. Hansen, P. R. (2005). A test for superior predictive ability. Journal of Business amp; Economic Statistics, 23(4), 365-380.
  4. Hsu, P., Hsu, Y., amp; Kuan, C. (2010). Testing the predictive ability of technical analysis using a new stepwise test without data snooping bias. Journal of Empirical Finance, 17(3), 471-484.

Why compare multiple forecasts to baseline forecast? Example

- Example for which you might want to do this
  - Look ar the performance across many fund managers to see if any (and who) beats the market return.
  - Examine the forecasts from a group of professional forecasters to see if any (and how many) produce exchange rate forecasts that beat the random walk.
  - Research team developed several forecasting models and wants to see which, and if any beats a random walk or AR(1) benchmark.

Why compare multiple forecasts to baseline forecast? Accounting for data snooping

- Accounting for data-snooping/ data-mining
  - Data snooping bias occurs when many models or forecasts are considered, but only the best are presented or published.
  - It can occur because a researcher only presents her best model.
  - It can occur hen many researchers examine the same data, but only the researchers with the most successful models get published (refer to as publication bias).
  - While abuses can be avoided some level of data mining is unavoidable part of research.

Data mining Cont.

- Data mining issues are also relevant in practice
  - A bank may hire many fund mangers but only advertise those that beat the market.
  - Newsletter scam (external & illegal example)
    - Each week 2 newsletters printed
    - One predicts market goes up next week
    - Other predicts market goes down next week
    - Sent to half of households
    - Following week, newsletters only sent to households that received correct prediction last week
    - Again, half get letters predicting that market falls
    - Week 3 - send only to households that received correct predictions in week 1 and 2.
    - It get repeated
    - ...
    - Pretty soon, there are a group of readers who think they have a newsletter that makes perfect predictions - they buy a subscription,

Framework

- e^0_{t+1} - baseline model forecast error
- e^m_{t+1} for $m = 1, ..., M$ forecast error from other models
- $L^0_{t+1} \equiv L(e^0_{t+1})$ - Loss from baseline forecast error
- $L^m_{t+1} \equiv L(e^m_{t+1})$ - Loss from model m
- $d^m_{t+1} = L^0_{t+1} - L^m_{t+1}$ - Loss from baseline froecast minus loss from model m.
- Model m produced a better forecast than "baseline" if $d^m_{t+1} > 0$ (m beat baseline at $t+1$).
- Model m produced better forecast on average if $\bar{d}^m > 0$ (m beat baseline on average)
  (Here $\bar{d}^m$ is average overforecast sample)
- Model m provides a better forecast than the baseline model if $\mu^m_d \equiv E[d^m_{t+1}] > 0$.

The basic multiple test problem

- Ask the question: do any of our forecasts beat the baseline models?

$$H_0: \ \mu_d^m \leq 0 \text{ all m (no model beats baseline)}$$
$$H_A: \ \mu_d^m > 0 \text{ for at least one m (Not } H_0\text{)}$$

- If M (number of forecast models small) we can conduct this joint test $\longrightarrow$ In which case there is no problem?
- But when M is large, this joint test is infeasible because:

$$\bar{d} = \begin{bmatrix} \bar{d}_1 \\ \dots \\ \bar{d}_M \end{bmatrix}$$

has $M \times M$ variance matrix, say $\underset{M \times M}{V}$ and joint test based on the wald statistics $W \equiv \bar{d}' \underset{M \times M}{\widehat{V}^{-1}} d$, which involves estimating and inverting the $M \times M$ variance-covariance matrix V.

The basic multiple test problem Cont.

- So realistically, we are forced to conduct multiple tests:

$$H_0^1 : \mu_d^1 \leq 0 \; H_A^1 : \mu_d^1 > 0 \text{ Reject if } t_1 = \frac{\bar{d}_1}{se(\bar{d}_1)} > Z_\alpha$$

$$H_0^2 : \mu_d^2 \leq 0 \; H_A^2 : \mu_d^2 > 0 \text{ Reject if } t_1 = \frac{\bar{d}_2}{se(\bar{d}_2)} > Z_\alpha$$

...

$$H_0^M : \mu_d^M \leq 0 \; H_A^M : \mu_d^M > 0 \text{ Reject if } t_1 = \frac{\bar{d}_M}{se(\bar{d}_M)} > Z_\alpha$$

- However, if significance level $\alpha = 0.05$ then we have designed the test so that about 5% of the tests reject by accident (Type I error) even if all the null hypothesis (Note: Since the tests statistics for the $m = 1, .., M$ are potentially corrected, the tests could reject by accident more or less than 5% but we use 5% example to keep discussion concrete).

The basic multiple test problem: A realistic solution

- For example, because there are so many funds, we are almost sure to find that some funds beat the market purely by accident.
- Similarly, data snooping bias refers to the fact if enough researchers try out enough models and enough variables some will show up significant purely by accident. And, these significant variables are the ones that will draw attention & lead to publication.
- The reality check solution in general terms
  - Instead of using M t-statistics, we focus on the largest t statistic
    Reject if $t_{max} = \underset{m=1,\ldots,M}{Max} t_m >$ Appropriate Critical Value
- Note that the model with largest t value shows most evidence against H_0.
- So if it isn't large enough to reject, then none of the t statistics.

How to pick the appropriate critical value?

- How to pick the appropriate critical value?

$$\bar{d} = \begin{bmatrix} \bar{d}_1 \\ \bar{d}_2 \\ \dots \\ \bar{d}_M \end{bmatrix} \sim N\left(\underset{M\times 1}{\mu_d}, \underset{M\times M}{V} \right)$$

Impose the equality pact of the null hypothesis. $H_o : \mu'_d = 0$
$\bar{d} \sim N(0, V)$. Then, by standardizing, we have

$$t = \begin{bmatrix} \bar{d}_1 / se_1 \\ \bar{d}_2 / se_2 \\ \dots \\ \bar{d}_M / se_M \end{bmatrix} = \begin{bmatrix} \bar{d}_1 / \sqrt{V_{11}} \\ \bar{d}_2 / \sqrt{V_{22}} \\ \dots \\ \bar{d}_M / \sqrt{V_{MM}} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{V_{11}}} & 0 & \dots & 0 \\ 0 & \frac{1}{\sqrt{V_{22}}} & \dots & 0 \\ \dots 0 & 0 & \dots & \frac{1}{\sqrt{V_{MM}}} \end{bmatrix} \begin{bmatrix} \bar{d}_1 \\ \bar{d}_2 \\ \dots \\ \bar{d}_M \end{bmatrix}.$$

How to pick the appropriate critical value? Cont.

- Call $\begin{bmatrix} \frac{1}{\sqrt{V_{11}}} & 0 & \dots & 0 \\ 0 & \frac{1}{\sqrt{V_{22}}} & \dots & 0 \\ \dots 0 & 0 & \dots & \frac{1}{\sqrt{V_{MM}}} \end{bmatrix} \equiv [diag(v)]^{-1/2}$.

$$t = diag(V)^{-1}\bar{d} \underset{H_0}{\sim} diag(V)^{-1/2}N(0, V) \sim N(0, diag(V)^{-1/2}Vdiag(V)^{-1/2})$$

Note that $diag(V)^{-1/2}Vdiag(V)^{-1/2}$ will have ones along diagonal, but it does not necessarily have zeros off diagonal $\longrightarrow$ not identity matrix.

- Therefore,

$$t_{max} = \underset{m-1,\dots,M}{Max} t \underset{H_0}{\sim} Z_{max}(V) = max\ N(0, diag(V)^{-1/2}Vdiag(V)^{-1/2}).$$

The distribution of the maximum of variables drawn from the multivariate normal distribution.

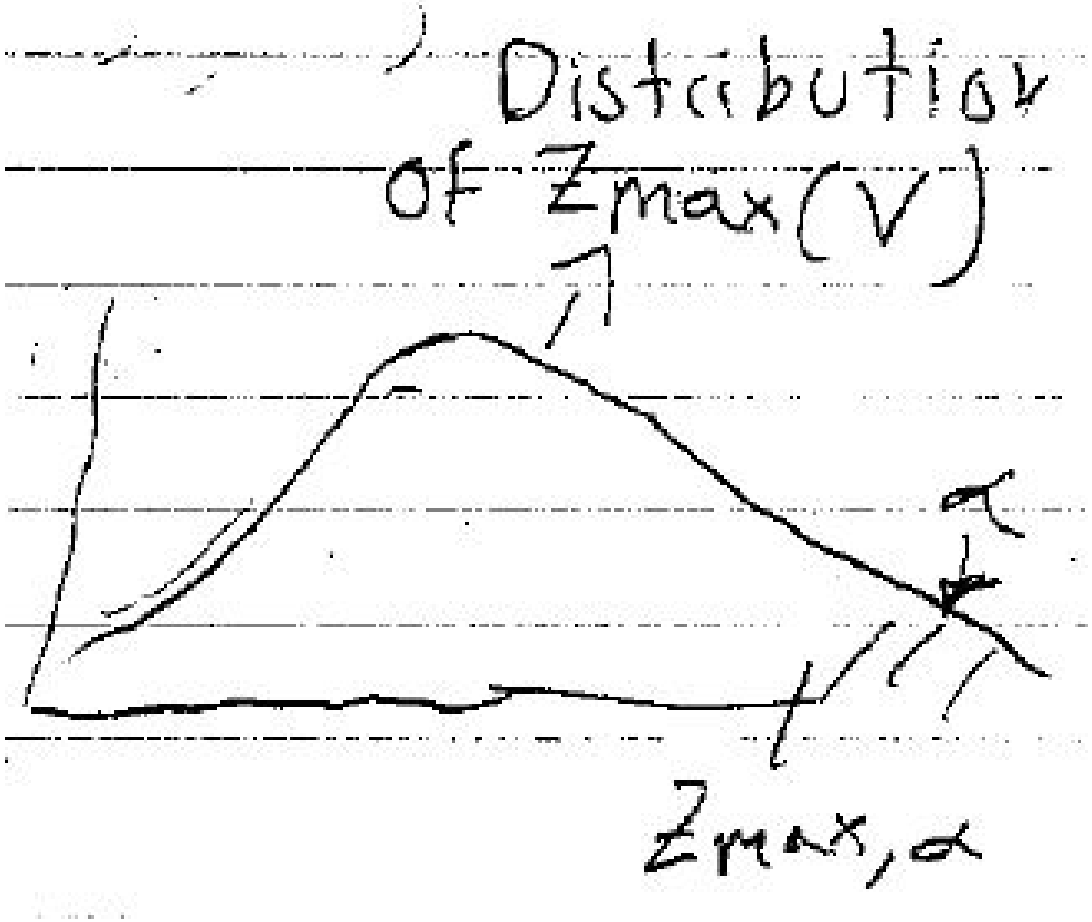
- Pick critical value (say $Z_{max,\alpha}$) such that $P(t_{max} > Z_{max,\alpha}) = \alpha$ Graph

Additional problems in finding $Z_{max,\alpha}$

- $Z_{max}(V)$ is Not a normal distribution
  - It is the maximum value of a draw from a normal distributions and the max of a normal RV is NOT a normal RV-cannot use Z lookup table.
  - $Z_{max}(V)$ depends on V and will be different in each application $\longrightarrow$ Impossible to provide alternative lookup table.
- Critical value is obtained by bootstrap resampling, which is itself a complicated procedure.
- Given enough time Ph.D. student could implement one of the four versions of the reality type test if needed for thesis chapter.
- Possibly, this could be done over summer for MA paper.
- For purposes of class project only feasible implementation would be using code available online or in software packages.

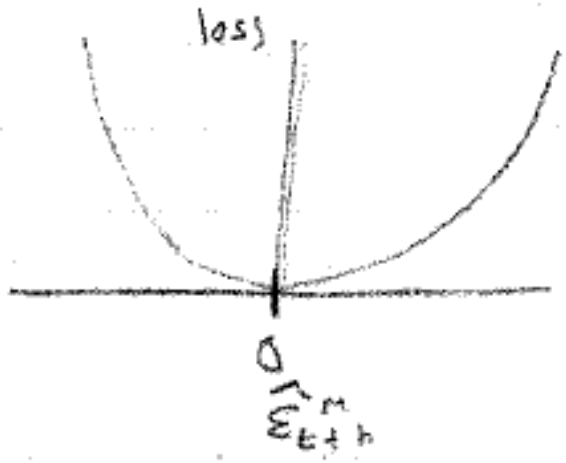
Procedures to select from

- White (2000) Reality Check Test
  1. original test
  2. tests if best model beats benchmark
- Hansen (2005) Superior Predictive Test (SPA test)
  1. improves power of white test
  2. test if best model beats forecast
- Romano & Wolf (2005) Stepwise Reality Test
  1. repeated reality check type tests to test for all models that beat the benchmark (not just the best)
  2. also some power improvements
- Hsu, Hsu & Kuan (2010) Stepwise Superior Predictive Test
  1. combine power improvements of Hansen with stepwise approach of Romano Wolf.

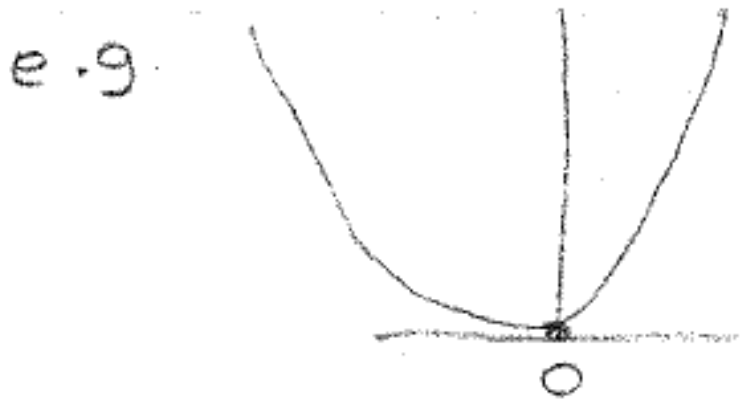


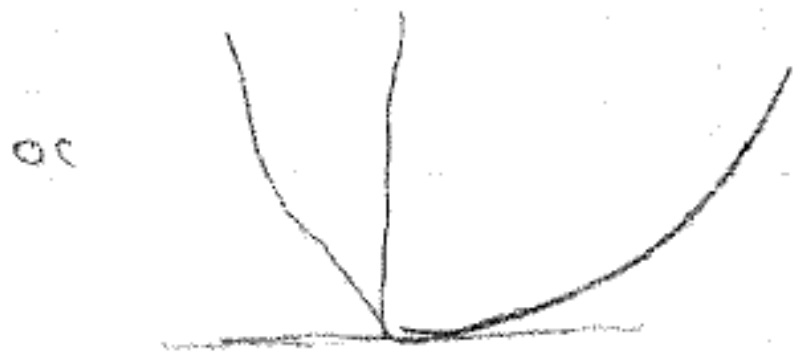
Back

e.g.

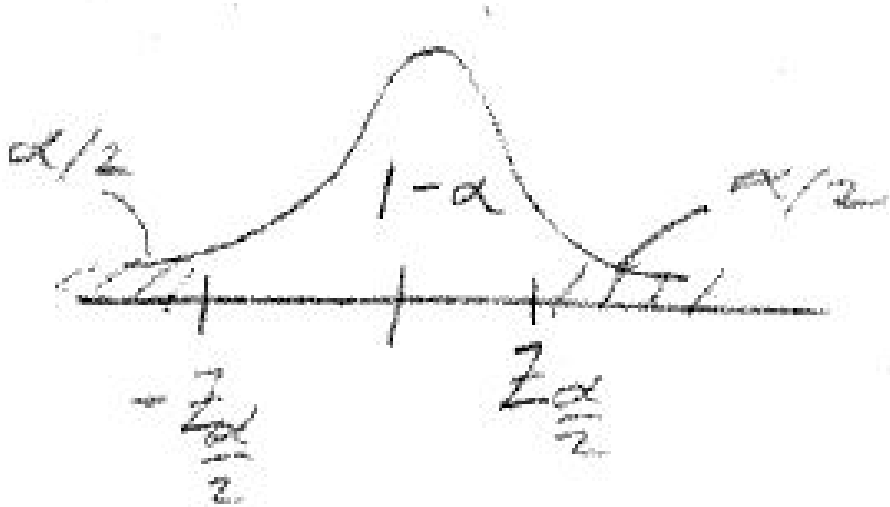


Back

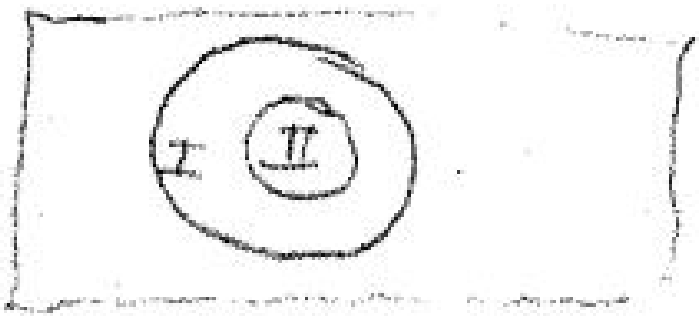




Back



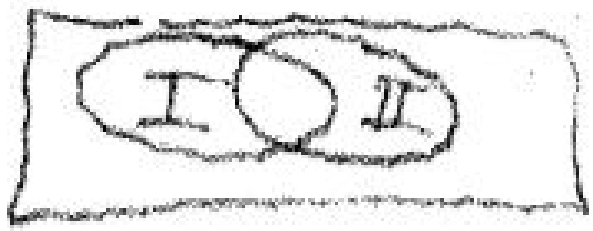
Back



I = info in model 1 forecast

II = info in model 2 forecast

Back



Back