Forecast comparison with macroeconomic models

Kevin Kotzé

Contents

Introduction

- We now consider the procedures for forecast estimation and evaluation
- These are relevant for both DSGE and ML models
- During these sessions we will place emphasis on the econometric foundations
- Also consider the empirical implications of using different test statistics
- The interested reader may enjoy the book by Elliot & Timmerman (2016)

Look at the data

- Before doing any forecasting the first thing that one would want to do is plot the data
- Look for changes in expected mean value of the underlying data-generating process
- Consider the degree of variability and potential changes in the variability
- One should also try to detect obvious outliers

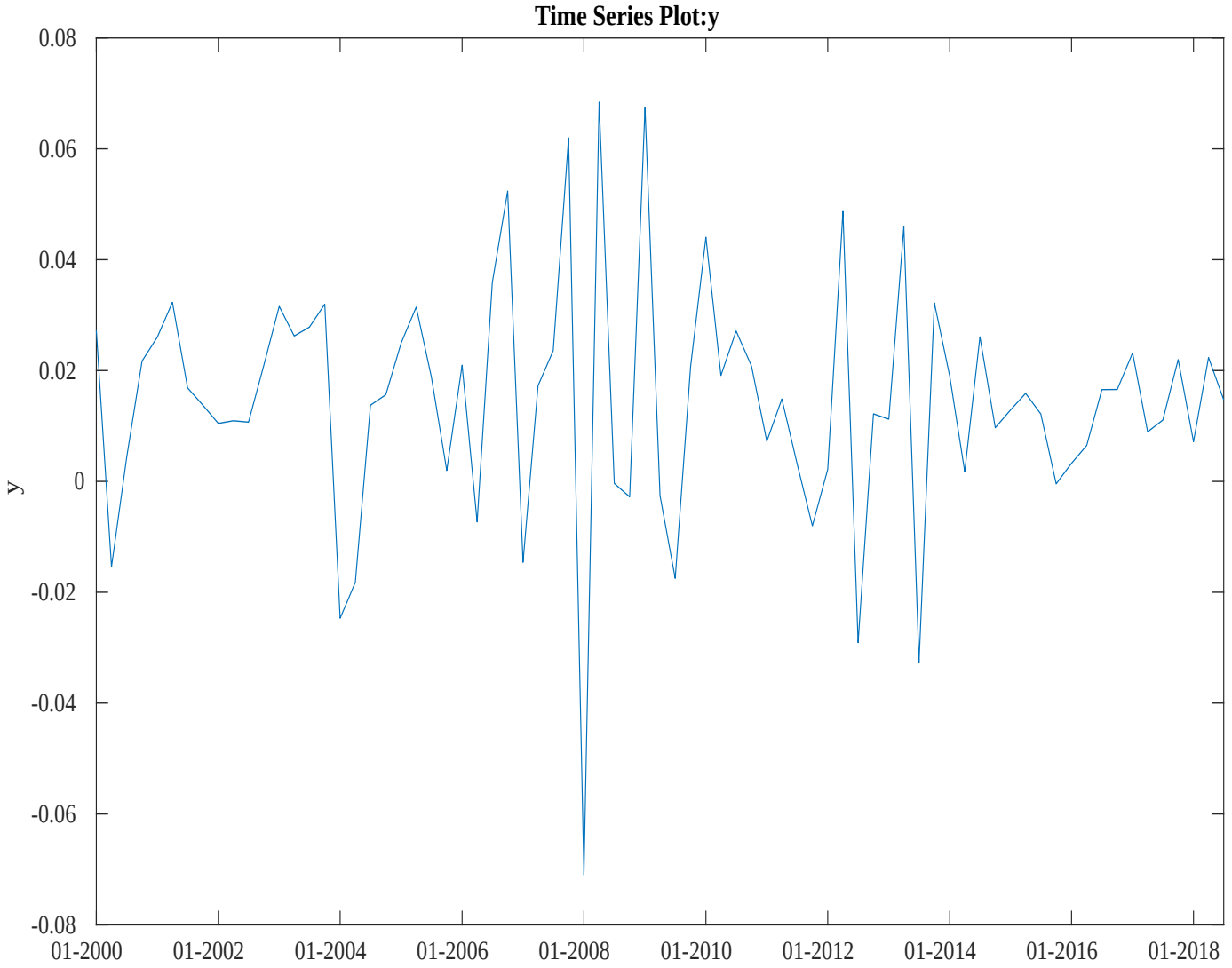

Figure - Output

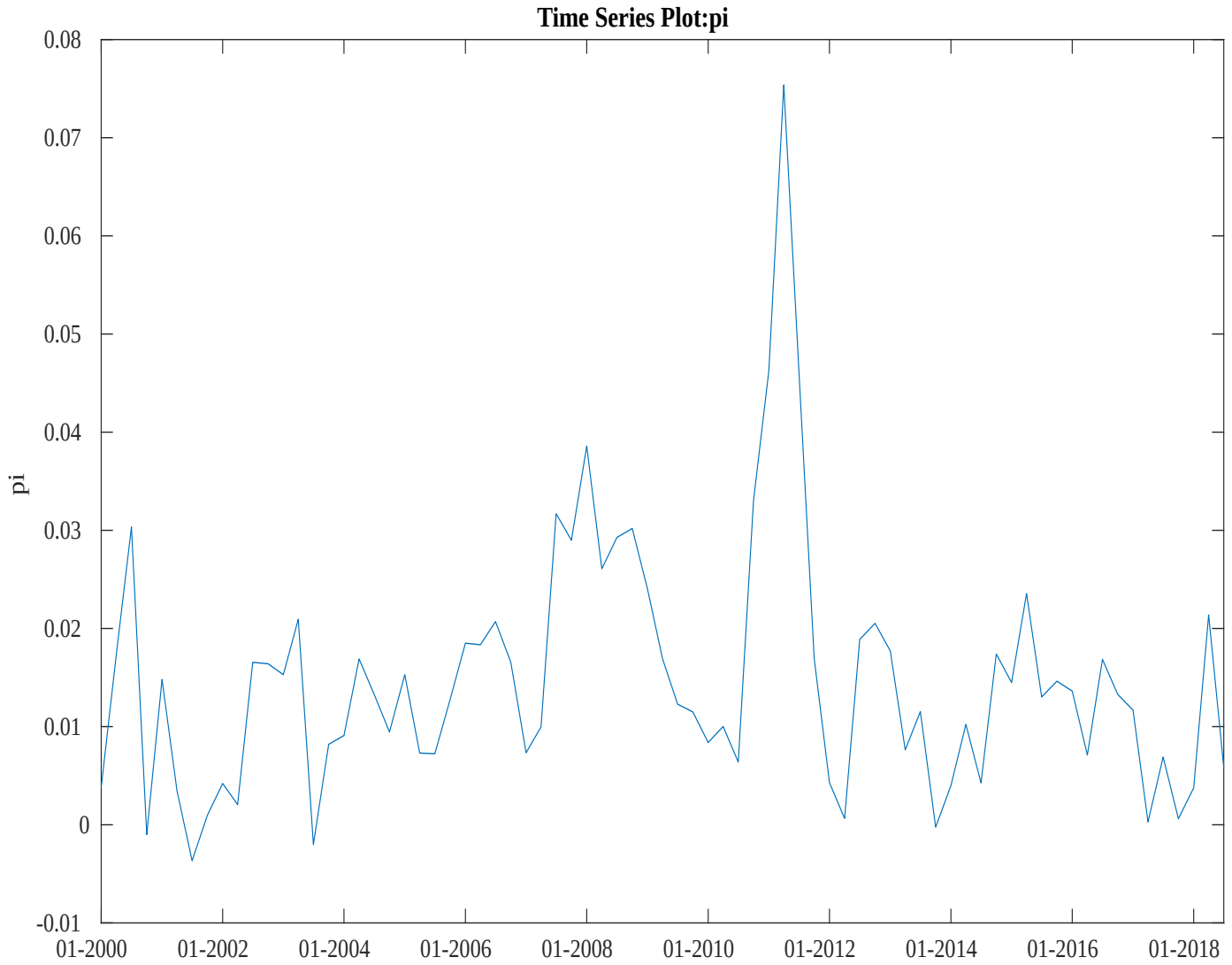

Figure - Inflation

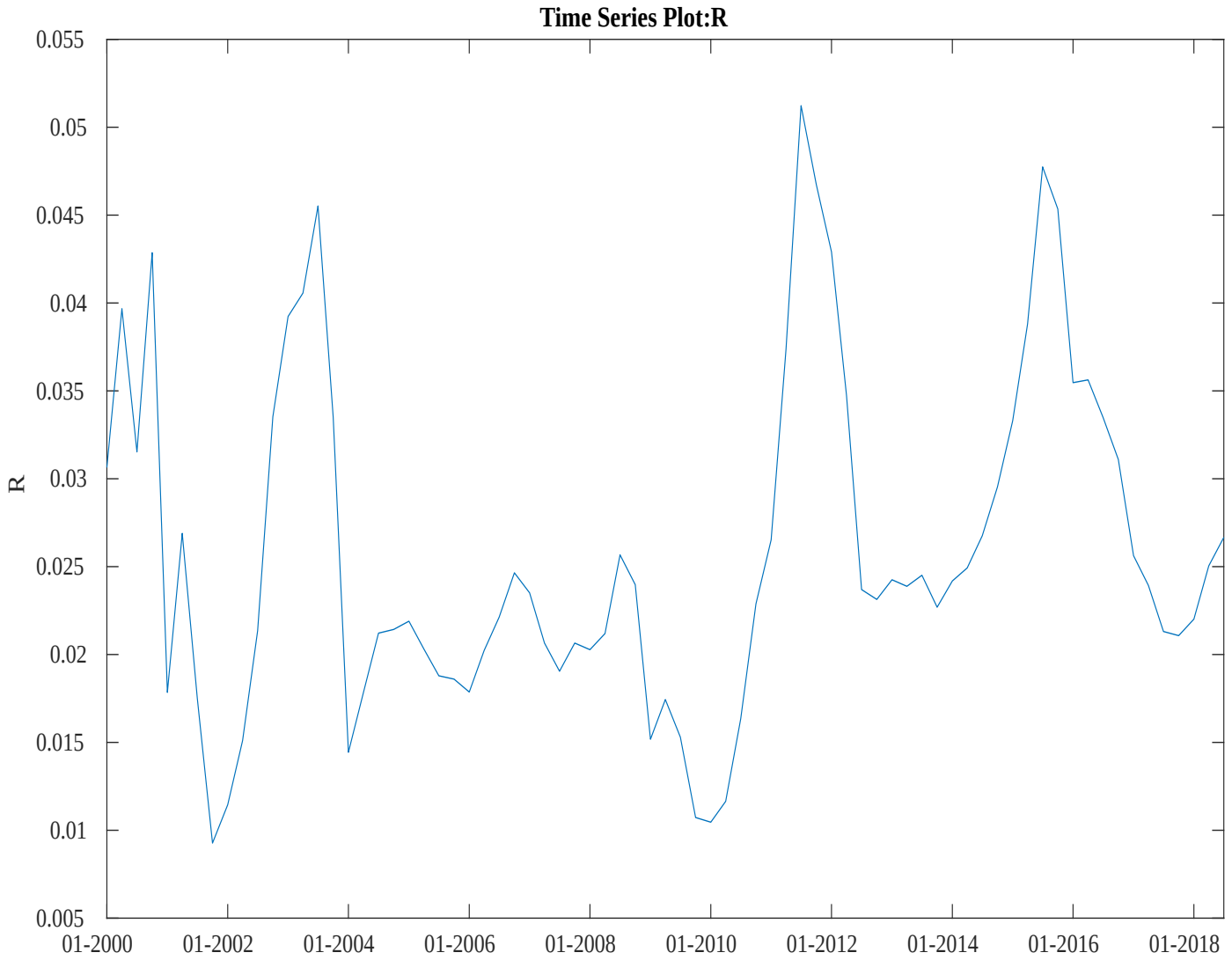

Figure - Interest rate

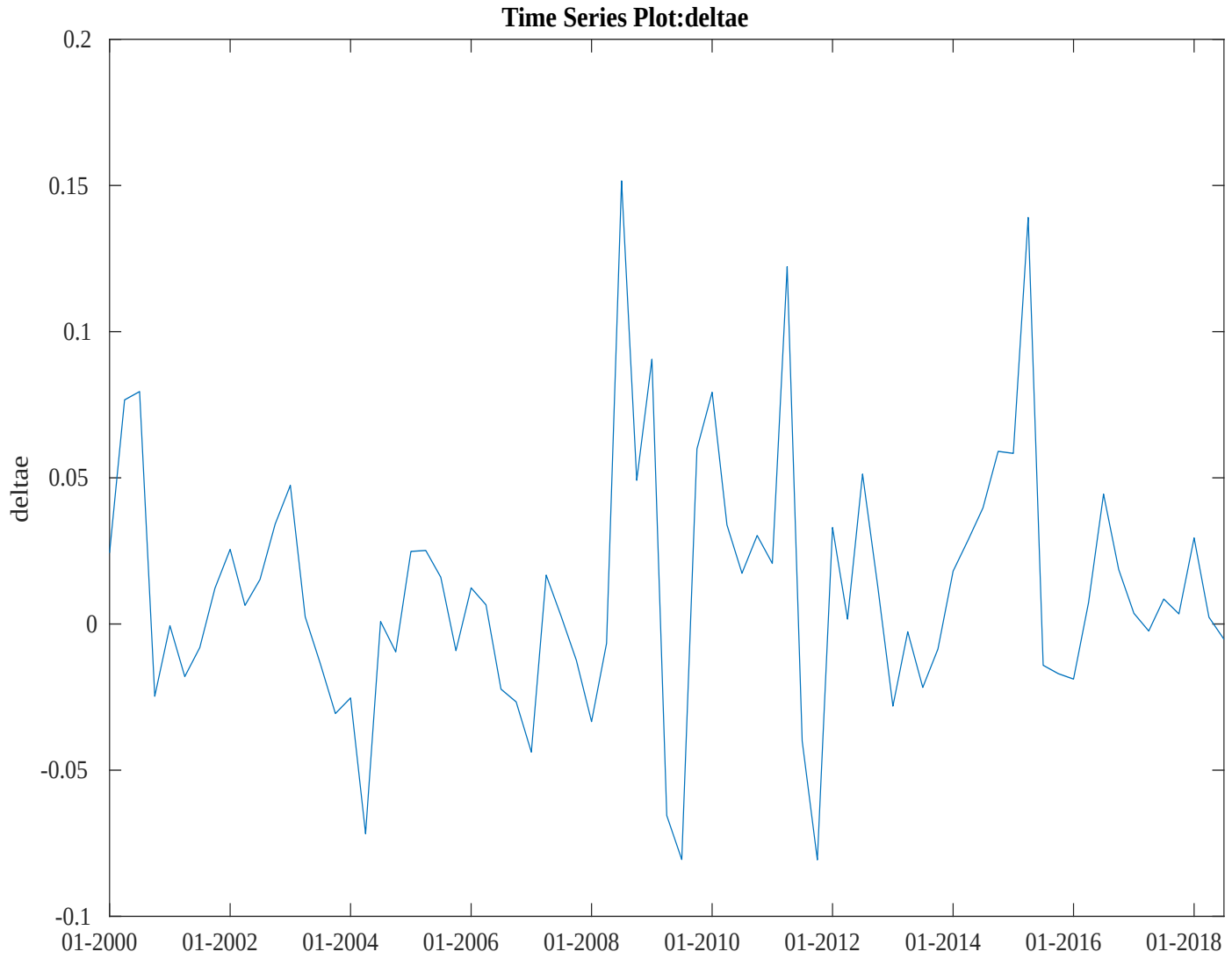

Figure - Exchange rate

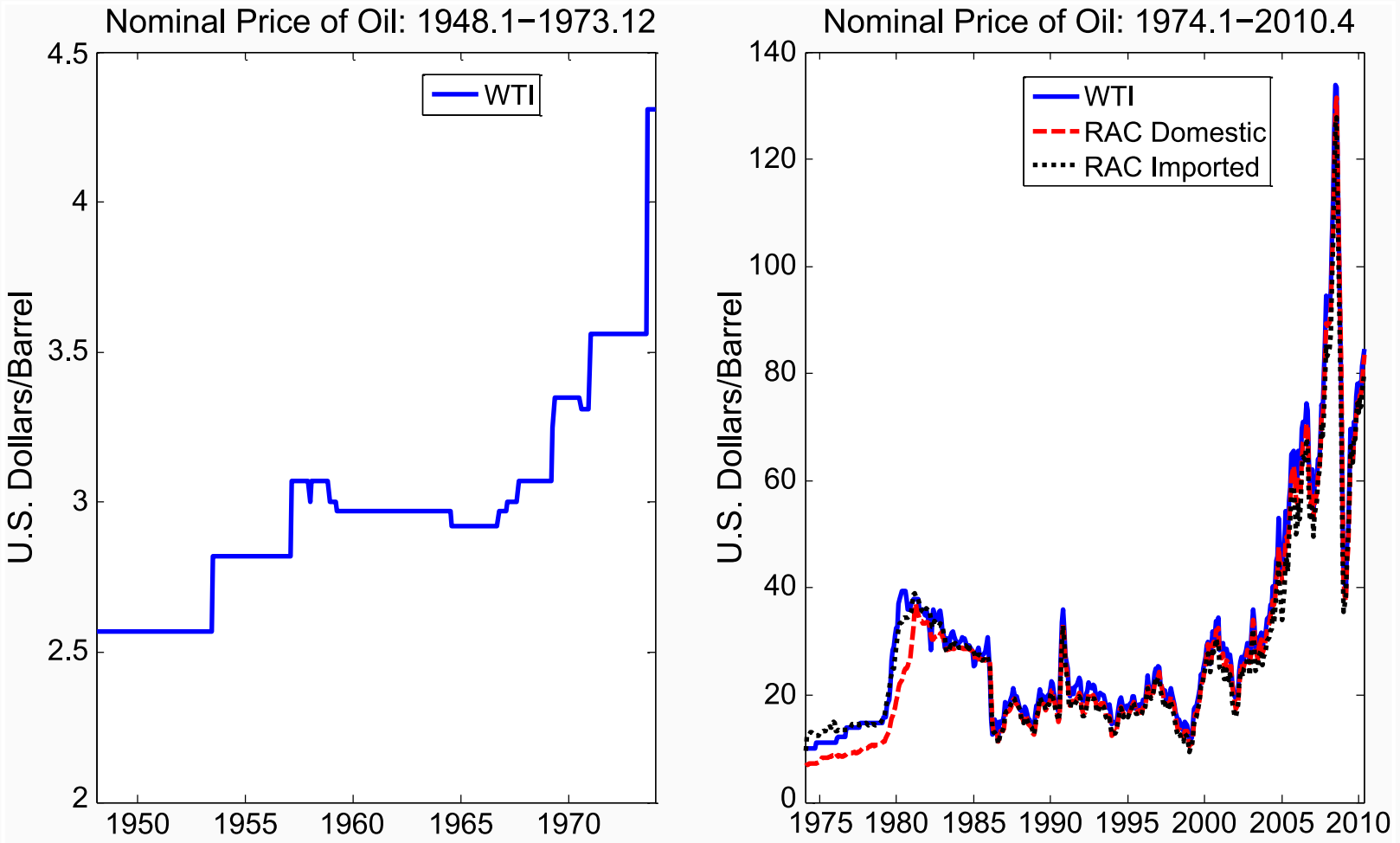

Figure 8.1 The nominal price of crude oil. *Notes*: WTI stands for the West Texas Intermediate price of crude oil and RAC for the U.S. refiners' acquisition cost.

Figure - Oil prices have a clear structural break

Notation

- Want to consider the use of various variables / models to forecast future values of a target variable
- Denote the target variable y_t and the set of predictors is x_t
- When we are at time t and we want to forecast h steps into the future
- Hence, we want to generate values for $\mathbb{E}_t\left[y_{t+h}\right]$ given some predictors
- Where there is linear relationship between these variables we use of the regression model:

$$\mathbb{E}_t\left[y_{t+h}\right] = \hat{\beta}^\top x_t$$

- where $\hat{\beta}$ would then represent a vector of estimated coefficients

Granger causality

- To consider if these predictors are any good we perform a Granger causality test
- This would tell us if $\hat{\beta}$ is significantly different from zero
- Construct the null for no predictive ability

$$H_0 : \beta = 0$$
$$H_1 : \beta \neq 0$$

- To perform the test we need to compare the estimated value of $\hat{\beta}$ against the restricted value under the null hypothesis
- Therefore, we are interested in the difference, $(\hat{\beta} - 0)$

Granger causality

- In addition we are also interested in the potential variability in the underlying data and associated estimate for β
- If there is large variability in the underlying data then we may conclude that larger value of $(\hat{\beta} - 0)$ may be insignificantly different from zero
- Therefore, when we only have a single predictor we could use a t-test:

$$t_\beta = \frac{\left(\hat{\beta} - 0\right)}{sd(\hat{\beta})}$$

- We reject the null hypothesis when the value for the t-test is large
- To reject the null within a 95% confidence interval, $|t_\beta| \geq 1.96$
- Alternatively, we could consider the p-values of the t-test

Granger causality

- Note that this statistic is conditional on the value of h
- A model could be useful when predicting one-step ahead, but useless at twelve-steps ahead
- When when we have several predictors we would need to construct an F-test
- To calculate an appropriate value for the denominator in the t-test we make use of the error term in the regression model

$$\mathbb{E}_t\left[y_{t+h}\right] = \beta^\top x_t + \epsilon_{t+h}, \quad \text{where} \quad \epsilon_{t+h} \sim \text{i. i. d.} \mathcal{N}(0, 1)$$

Variance of the denominator

- When using lags of the dependent variable to forecast forward the shocks to the variable are not independent
- Consider the AR(1) that may be used for successive h step-ahead forecasts:

$$
\begin{aligned}
\mathbb{E}_t\left[y_{t+1}\right] &= \rho y_t + \epsilon_{t+1} \\
\mathbb{E}_t\left[y_{t+2}\right] &= \rho y_{t+1} + \epsilon_{t+2} = \rho\left(\rho y_t + \epsilon_{t+1}\right) + \epsilon_{t+2} \\
\mathbb{E}_t\left[y_{t+3}\right] &= \rho y_{t+2} + \epsilon_{t+3} = \rho\left(\rho y_{t+1} + \epsilon_{t+2}\right) + \epsilon_{t+3} \\
&\vdots = \vdots
\end{aligned}
$$

- where the values for ϵ_{t+h} represent the forecast errors
- The larger is h the more serially correlated the error term
- Need to use Newey & West (1987) heteroskedasticity and autocorrelation consistent (HAC) estimate of the variance

Critique of Granger causality

- If x_t Granger causes y_t, it would not necessarily be a useful predictor of y_t
- Messe & Rogoff (1983) note that although the interest rate differential $i_{t+h} - i^*_{t+h}$ Granger causes a change in the exchange rate, $s_{t+h} - s_t$, it does not necessarily perform well when used in an out-of-sample forecast evaluation exercise
- Part of the intuition behind this result is that impressive in-sample Granger causality statistics may be due to over-fitting or structural breaks
- This promoted the use of out-of-sample forecasting exercises

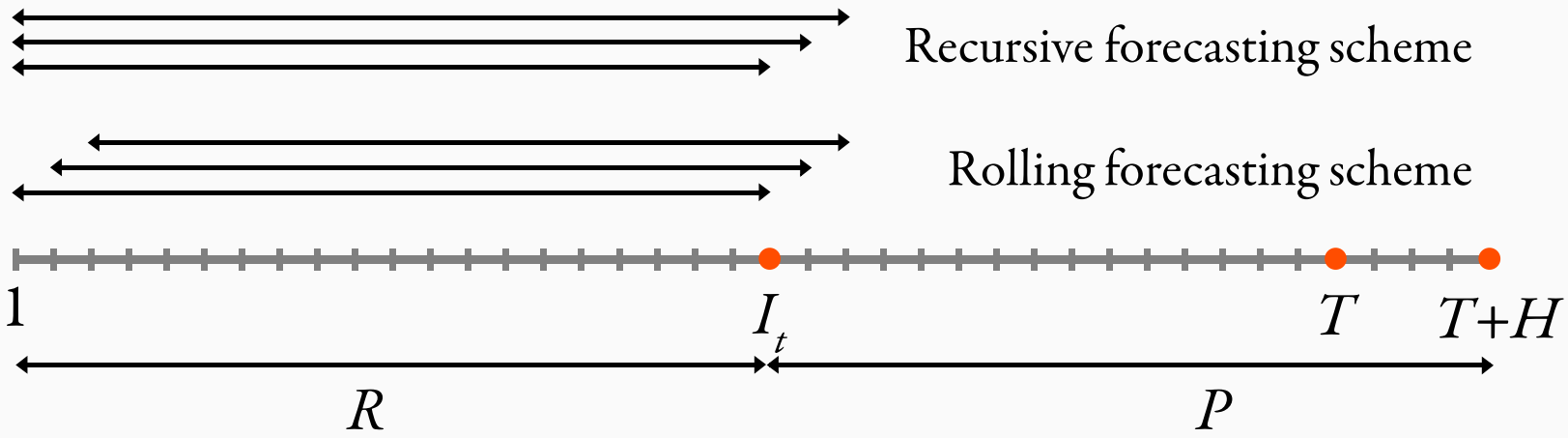

Figure - Out-of-sample notation

Out-of-sample forecasting

- Make use of an in-sample portion of size R (i.e. training dataset)
- Out-of-sample portion of size P, which refers to all the predictions (i.e. test dataset)
- To generate the first forecast for $\mathbb{E}_t \left[y_{t+1} \right]$ we would make use of the data until period R to generate a forecast for period $R + 1$
- The in-sample period used in the estimation may be termed the *information set*, denoted I_t
- If we have quarterly data and want to obtain forecasts over the next eight quarters (i.e. two years), then we would want to generate eight successive forecasts, where $h = \{1, 2, \ldots, 8\}$
- The end of the forecasting horizon may be represented by H, where in this case, $H = 8$

Out-of-sample evaluation

- After generating the forecast y^f_{R+h} we can compare it to the realised value of this time series y_{R+h}
- when comparing these values we generate the forecast error

$$\varepsilon^f_{R+h} = y_{R+h} - y^f_{R+h}$$

- To evaluate the accuracy of the forecast we mimic what the forecaster would have done in real time
- This involves estimating the model over various points of time to generate a sequence of forecast errors

Out-of-sample evaluation

- After estimating the model with the use of data for the in-sample period R to calculate the initial forecast errors, the in-sample period would then increase to $R + 1$
- Model is then re-estimated to generate a new value set of $\hat{\beta}$ coefficients to generate the new set of forecasts for the period y^f_{R+h+1}
- This procedure would continue until the last estimation, which takes place at time T to generate forecasts up until period $T + H$

Out-of-sample evaluation

- With a recursive scheme, the initial observation in the in-sample period is fixed at the first observation
- For a rolling-window scheme we maintain a fixed number of observations for the in-sample period
- Hence, the rolling-window scheme would be preferred when there are potential structural breaks in the in-sample period
- Recursive scheme may produce more accurate forecasts when the target variable is relatively stable
- When comparing the forecasts to the realised values of economic data, which may be revised by the statistical office
- This promted individuals to store different vintages of data following the first period release of such data

Out-of-sample evaluation

- This procedure provides a sequence of forecast errors that may be expressed as follows:

$$\left\{\varepsilon_{t+h}^{f}\right\}_{t=R}^{T}$$

- Therefore we have a total of P forecast errors for each horizon, h
- Could compare forecasts to real time vintages of data to exclude the effects of data revisions
- Mimic what the forecaster would have done over a period of time to generate a sequence of forecast errors

Model comparison

- To consider if these forecasts are any good we could compare the results against a random-walk model
- Such a model may be expressed as:

$$\mathbb{E}_t \boldsymbol{y}_{t+h} = \boldsymbol{y}_t$$

- Such a model could then be used to generate a second series of forecast errors:

$$\left\{ \varepsilon_{t+h}^{rw} \right\}_{t=R}^{T}$$

Model comparison

- To measure whether the forecast errors are centred around zero we can take the simple average of the forecast errors
- This provides an estimate of the forecast bias:

$$\sum_{t=R}^{T} \left\{ \varepsilon_{t+h}^{f} \right\} \quad \text{and} \quad \sum_{t=R}^{T} \left\{ \varepsilon_{t+h}^{rw} \right\}$$

- This would not be the only statistic of interest as a model that has very large over-predictions along with very large under-predictions could still provide a bias of zero
- We need to ensure that these positive and negative errors do not necessarily cancel each other out

Model comparison

- Also need give due consideration to the particular loss function that would be relevant to problem at hand
- In certain cases we may be concerned by over-prediction more than what we are concerned about under-prediction
- This would call for the use of an asymmetric loss function
- Also want to consider the degree to which we are concerned by outliers in the forecasting error as this would also influence our choice of loss function

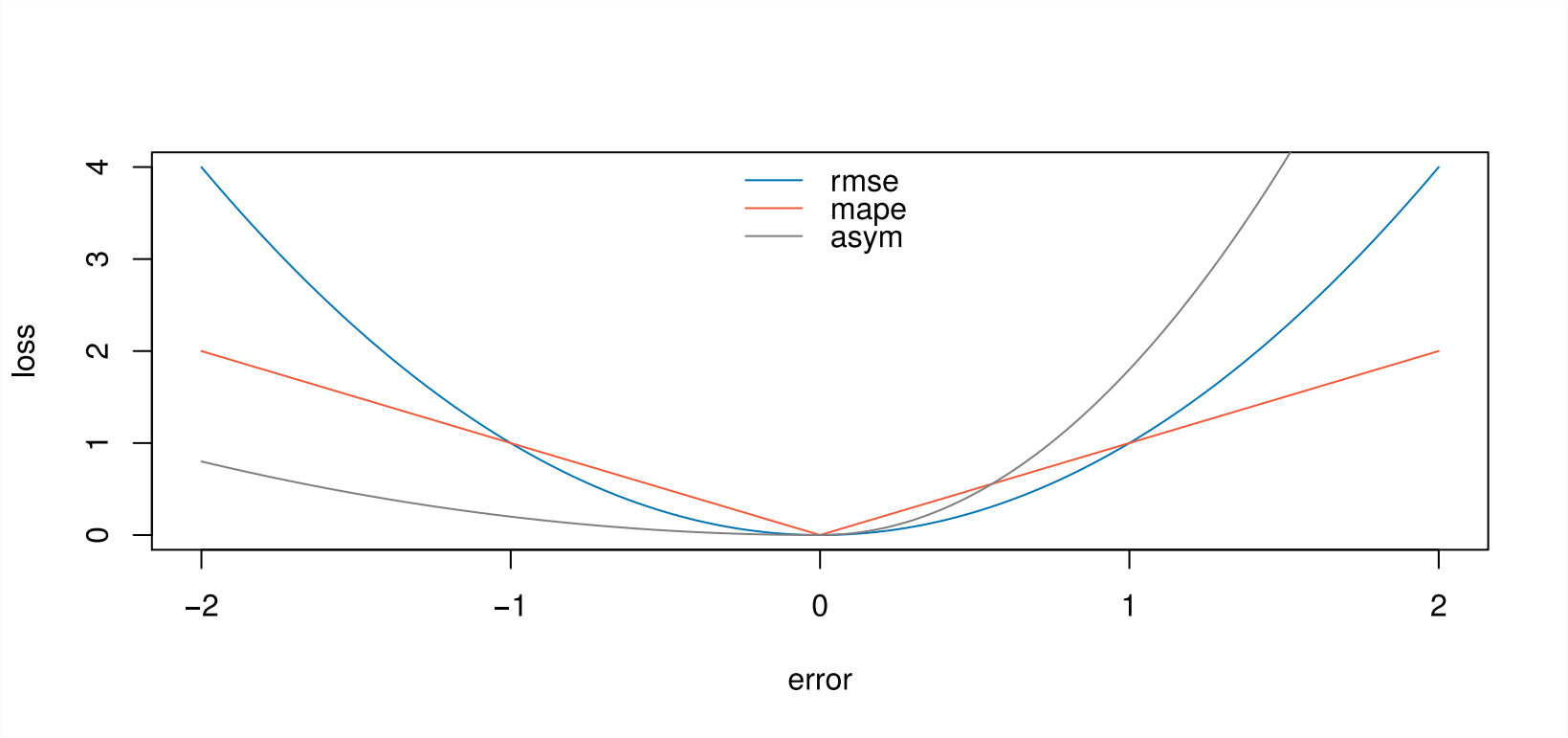

Figure - Loss functions

Model comparison

- Most popular of these three is a squared loss function that is used in the root-mean squared error (RMSE)
- Work with the mean squared error (MSE) which is a quadratic function that places a larger penalty on large errors
- The mean average predictive error (MAPE) makes use of the absolute value of the forecasting error, which provides a linear loss function
- In the literature there are many different functional forms for loss functions including various asymmetric loss functions

Model comparison

- We denote the loss function $\mathscr{L}_{t+h}$ for the forecast that is h steps-ahead
- The loss function could then be used to evaluate the forecasts from competing models, after we have generated the sequence of forecast errors
- In the above example we have two sequences of forecast errors for $\left\{\varepsilon^{f}_{t+h}\right\}^{T}_{t=R}$ and $\left\{\varepsilon^{rw}_{t+h}\right\}^{T}_{t=R}$

Model comparison

- To identify the model that produces the most accurate prediction we use a quadratic loss function in what follows
- We firstly take of the sum of the square of the individual forecasts errors from each model before calculating the difference:

$$\Delta\mathscr{L}_{t+h} = \sum_{t=R}^{T} \left\{\varepsilon_{t+h}^{f}\right\}^{2} - \sum_{t=R}^{T} \left\{\varepsilon_{t+h}^{rw}\right\}^{2}$$

- If the result is positive then it would suggest that the model that was used to generate ε_{t+h}^{f} is inferior
- However, if this difference is negligible then we may want to suggest that the models have equal predictive ability

Model comparison

- To formally test whether a model has equal predictive ability we make use of a t-test
- The null hypothesis is that the expected difference in the forecast errors is not different from zero

$$
\begin{aligned}
H_0 \ &: \ \mathbb{E}_{T+H}\left[\sum_{t=R}^{T}\left\{\varepsilon_{t+h}^{f}\right\}^2 - \sum_{t=R}^{T}\left\{\varepsilon_{t+h}^{rw}\right\}^2\right] = 0 \\
\therefore \ H_0 \ &: \ \mathbb{E}_{T+H}\left[\Delta\mathscr{L}_{t+h}\right] = 0
\end{aligned}
$$

Model comparison

- To perform this test we can take the sequence of values for $\Delta\mathscr{L}_{t+h}$ and regress it on a constant, c
- In this case the $\hat{\beta}$ coefficient could be subjected to a t-test to see if it is significantly different from zero in the regression:

$$\Delta\mathscr{L}_{t+h} = \hat{\beta}c + \epsilon_{t+h}$$

Model comparison

- And the t-test would be expressed as:

$$t_\beta = \frac{\hat{\beta} - 0}{sd\left(\hat{\beta}\right)}$$

- Use a HAC estimate of the standard deviation of $\Delta\mathscr{L}_{t+h}$
- In this case the estimated value of $\hat{\beta}$ would be equal to the average value of the difference in the loss function

$$\hat{\beta} = \frac{1}{P}\sum_{t=R}^{T}\Delta\mathscr{L}_{t+h} = \Delta\bar{\mathscr{L}}_{t+h}$$

- Such that

$$t_\beta = \frac{\Delta\bar{\mathscr{L}}_{t+h} - 0}{sd\left(\Delta\bar{\mathscr{L}}_{t+h}\right)}$$

Diebold & Mariano test

- We consider whether the $\hat{\beta}$ coefficient is different from zero with a t-test

$$t_{DM} = \frac{\hat{\beta} - 0}{sd\left(\Delta\bar{\mathscr{L}}_{t+h}\right)} = \frac{\hat{\beta} - 0}{\sqrt{var\left(\Delta\bar{\mathscr{L}}_{t+h}\right)}} = \frac{\hat{\beta} - 0}{\sqrt{var\left(\frac{1}{P}\sum_{t=R}^{T}\Delta\mathscr{L}_{t+h}\right)}}$$

$$\therefore t_{DM} = \frac{\hat{\beta} - 0}{\sqrt{var\left(\sum_{t=R}^{T}\Delta\mathscr{L}_{t+h}\right)}}\sqrt{P}$$

- Due to correlation in the residual we use a HAC estimate of the standard deviation of the $\Delta\mathscr{L}_{t+h}$ sequence
- To evaluate the result we conclude that the models do not have equal forecasting ability when the t-statistic is large

Diebold & Mariano test

- Diebold & Mariano (1995) show that the distribution of the difference in the loss function converges to the distribution of the normal distribution when the number of forecasts that we have is sufficiently large (i.e. more than 100 observations)
- Implies that when $|t_{DM}| \leq 1.96$ we are unable to reject the null when working with a 95% confidence interval
- In essence these authors are responsible for the development of a new literature on how to compare the predictive ability of different models

Parameter estimation error

- Significant limitation of Diebold & Mariano (1995) test is that we frequently use an estimate for $\hat{\beta}$ to generate predictions that are not actually observed
- Other models, such as the random-walk do not make use of parameter estimates
- Typically the variability of the loss that is based on a parameter estimate will be greater than the loss that is not based on a parameter estimate
- Reason for this is that the loss that is calculated from the linear regression is:

$$\epsilon_{t+h} = y_{t+h} - \hat{\beta}^{\top} x_t$$

- This would be more variable if there is a large degree of parameter uncertainty relating to the value of $\hat{\beta}$
- Hence, such a model would usually under-estimate the necessary degree of variability of the losses

Parameter estimation error

- The effect of the estimation error on the degree of variability in the forecast error is considered in West (1996)
- Compares the forecasting errors for one model that includes estimations errors, which is denoted below by M_1, and another that does not

$$M_1 : \mathscr{L}_T^{\hat{\beta}} = \frac{1}{P} \sum_{t=R}^{T} \left\{ y_{t+h} - \hat{\beta}^\top x_t \right\}^2$$

$$M_2 : \mathscr{L}_T^{\beta} = \frac{1}{P} \sum_{t=R}^{T} \left\{ y_{t+h} - \beta^\top x_t \right\}^2$$

Parameter estimation error

- To consider the properties of the distribution of $\hat{\beta}$ he makes use of a mean value expansion of $\hat{\beta}$ around β
- Such an expansion could take the form

$$\Delta \bar{\mathscr{L}}_T^{\hat{\beta}} \approx \Delta \bar{\mathscr{L}}_T^{\beta} + \frac{\partial \bar{\mathscr{L}}_T^{\beta}}{\partial (\beta)} \left[\left(\hat{\beta} - \beta \right) \sqrt{R} \right] \sqrt{\frac{P}{R}}$$

- where $\frac{\partial \bar{\mathscr{L}}_T^{\beta}}{\partial (\beta)} \left[\left(\hat{\beta} - \beta \right) \sqrt{R} \right] \sqrt{\frac{P}{R}}$ is the component that is due to the estimation error

Parameter estimation error

- Therefore, in such cases the Diebold & Mariano (1995) test would need to be amended, such that

$$t_W = \frac{\hat{\beta} - 0}{\sqrt{var\left(\sum_{t=R}^{T} \Delta\mathscr{L}_{t+h}\right) + z_t}}\sqrt{P}$$

- Where z_t is the contribution of parameter estimation error that is due to the variance of the losses
- This allowed West (1996) to show that if the estimation error decreases when the in-sample period increases, then the effect of parameter estimation decreases when R increases for fixed values of P

Nested models

- When models are nested then some of the coefficients from the larger model are not available in the restricted model
- When constructing the null hypothesis, the variance for $\Delta\bar{\mathscr{L}}_T^{\hat{\beta}}$ would be equal to zero

$$\Delta\bar{\mathscr{L}}_T^{\hat{\beta}} = \left(y_{t+h} - \hat{\beta}^\top x_t\right)^2 - \left(y_{t+h} - 0\right)^2 = 0 \;\; \forall t$$

- Therefore when $\hat{\beta} = 0$ then the variance will be zero and we are not able to use the Diebold & Mariano (1995) test
- This limitation is discussed in Clark & McCracken (2001) and it is an important case, as the random walk model is often nested within other models, or where a VAR model is nested in a DSGE model

Nested models

- They propose the use of an encompassing test that could be expressed as follows:

$$ENCNEW = P\frac{\frac{1}{P}\sum_{t=R}^{T}\left(\epsilon_{1,t+h}^{2} - \epsilon_{1,t+h}\epsilon_{2,t+h}\right)}{\frac{1}{P}\sum_{t=R}^{T}\epsilon_{2,t+h}^{2}}$$

- Where $\epsilon_{1,t+h}$ are the forecast errors of the small model and $\epsilon_{2,t+h}$ are the forecast errors of the large model
- Note that as the denominator is not represented by the variance of $\Delta\mathscr{L}_T$, this statistic will not be subject to the same problems as the Diebold & Mariano (1995)

Nested models

- This statistic does not have a standard distribution and as such the critical values would need to be calculated from a Monte Carlo simulation for $h = 1$
- These critical values are included in their paper and should only be used for linear models
- To calculate values for a different $h = 1$, we would need to apply a bootstrap to the simulation

Nested models

- To generate a test statistic that could be compared to a normal distribution Clark & West (2007) compare the properties of nested model with the properties of a model that would be subject to a normal distribution
- In doing so they look to adjust the test statistic in a way that would allow it to be compared to a normal distribution
- So they consider the forecasts of a large model, where the forecasts are generated by $\mathbb{E}_t\left(y_{t+h}\right) = \hat{\beta}^\top x_t$ along with a small random-walk model that has forecasts $\mathbb{E}_t\left(y_{t+h}\right)$

Nested models

- In the case of both models:

$$MSFE_{large} = \frac{1}{P}\sum_{t=R}^{T}\left(y_{t+h} - \hat{\beta}^{\top} x_t\right)^2$$

$$MSFE_{small} = -\frac{1}{P}\sum_{t=R}^{T}\left(y_{t+h}\right)^2$$

- The essential idea is to adjust the mean-squared forecasting error of the large model to correct for the effects of parameter uncertainty when calculating the test statistic

Nested models

- If one were to apply the Diebold & Mariano (1995) test to this problem we would proceed as follows:

$$\begin{aligned}
\Delta\bar{\mathscr{L}}_T &= \frac{1}{P}\sum_{t=R}^{T}\left(y_{t+h} - \hat{\beta}^\top x_t\right)^2 - \frac{1}{P}\sum_{t=R}^{T}\left(y_{t+h}\right)^2 \\
&= \frac{1}{P}\sum_{t=R}^{T}\left(y_{t+h}\right)^2 + \frac{1}{P}\sum_{t=R}^{T}\left(\hat{\beta}^\top x_t\right)^2 - \frac{2}{P}\sum_{t=R}^{T}\left(y_{t+h}\right)\left(\hat{\beta}^\top x_t\right) - \dots \\
&\quad -\frac{1}{P}\sum_{t=R}^{T}\left(y_{t+h}\right)^2 \\
&= \frac{1}{P}\sum_{t=R}^{T}\left(\hat{\beta}^\top x_t\right)^2 - \frac{2}{P}\sum_{t=R}^{T}\left(y_{t+h}\right)\left(\hat{\beta}^\top x_t\right)
\end{aligned}$$

Nested models

- Now under the null hypothesis $y_{t+h} = \epsilon_{t+h}$ such that,

$$
\begin{aligned}
H_0 : \frac{1}{P}\sum_{t=R}^{T}\left(\hat{\beta}^{\top}x_t\right)^2 - \frac{2}{P}\sum_{t=R}^{T}\left(\epsilon_{t+h}\right)\left(\hat{\beta}^{\top}x_t\right) \\
: \frac{1}{P}\sum_{t=R}^{T}\left(\hat{\beta}^{\top}x_t\right)^2 - \frac{2}{P}\hat{\beta}\sum_{t=R}^{T}\left(\epsilon_{t+h}\right)\left(x_t\right)
\end{aligned}
$$

- Note that if we assume that ϵ_{t+h} are independent then $\frac{2}{P}\hat{\beta}\sum_{t=R}^{T}\epsilon_{t+h}x_t = 0$
- Since the value for $\frac{1}{P}\sum_{t=R}^{T}\left(\hat{\beta}^{\top}x_t\right)^2$ will always be positive the loss of the large model will always exceed the small model

Nested models

- This statistic is used in Clark & West (2007) to correct the Diebold & Mariano (1995) to correct for the unfair advantage of the smaller model
- Since the distribution for $\frac{1}{P}\sum_{t=R}^{T}\left(\hat{\beta}^{\top}x_t\right)^2$ should be normal, when correcting the the Diebold & Mariano (1995) test, which is asymptotically normal by the distribution of something that is also normally distributed, we are left with a test statistic that is asymptotically normal
- Therefore the Clark & West (2005) statistic applies a similar framework to Diebold & Mariano (1995) test, but where the measure of $\Delta\bar{\mathscr{L}}_T$ is adjusted for the estimation error

$$t_{CW} = \frac{\Delta\bar{\mathscr{L}}_T^{adj} - 0}{\sqrt{var\left(\bar{\mathscr{L}}_T^{adj}\right)}} \underset{H_0}{\longrightarrow} \mathcal{N}(0,1)$$

Nested models

- The utlimate effect of this is that it would be easier for the large model to outperform the small model when using the Clark & West (2007) as opposed to using the Diebold & Mariano (1995) test on nested models
- Note that this is a one-sided test, where we are only testing the null that the models have equal predictive ability, or that the large model provides more accurate estimates
- These statistics do not test whether or not the small model provides more accurate forecast estimates
- In addition all of these tests focus on the null hypothesis, $H_0 : \beta = 0$, based on $\mathbb{E}\left(\Delta\mathscr{L}_{t+h}^{\beta}\right) = 0$, where β is the true parameter estimate
- To make use of a different null hypothesis, $H_0 : \mathbb{E}\left(\Delta\mathscr{L}_{t+h}^{\hat{\beta}}\right) = 0$, where $\hat{\beta}$ is the sample estimate of the population parameter, consult the work of Giacomini & White (2006)

Summary

- In the case of the out-of-sample tests the value of the coefficients are changing with each successive estimation
- This would not be the case in the in-sample Granger causality tests
- In the next few sessions we are going to focus on the use and limitations of some of these evaluation tests