Lecture 15
Forecasting

RS (for private use, not to be posted/shared online). 1

Forecasting

• A shock is often used to describe an unexpected change in a variable or in the value of the error terms at a particular time period.

• A shock is defined as the difference between expected (a forecast) and what actually happened.

• One of the most important objectives in time series analysis is to forecast its future values. It is the primary objective of ARIMA modeling:

• Two types of forecasts.
- **In sample** (prediction): The expected value of the RV (in-sample), given the estimates of the parameters.
- **Out of sample** (forecasting): The value of a future RV that is not observed by the sample.

For private use only - Not to be shared/posted online without authorization.

ARIMA: Forecasting

• Forecasting is the primary objective of ARIMA modeling.

• Two types of forecasts.
- **In sample** (prediction): The expected value of the RV (in-sample), the "fitted values," $\hat{Y}_t$.
- **Out of sample** (forecasting): The value of a future RV that is not observed by the sample, $\hat{Y}_{T+\ell}$. This is what we are going to do.

• Forecast: Conditional expectation of $Y_{T+\ell}$, given I_T:

$$\hat{Y}_{T+\ell} = E[Y_{T+\ell} \,|\, I_T = \{Y_T, Y_{T-1}, \ldots, Y_1, \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_T\}]$$

Notation:
- Forecast for $T+\ell$ made at T: $\hat{Y}_{T+\ell}$, $\hat{Y}_{T+\ell|T}$, $\hat{Y}_T(\ell)$.
- $T+\ell$ forecast error: $e_{T+\ell} = e_T(\ell) = Y_{T+\ell} - \hat{Y}_{T+\ell}$

ARIMA: Forecasting – Basic Concepts

• The variable to forecast $Y_{T+\ell}$ is a RV. It can be fully characterized by a pdf.

• In general, it is difficult to get the pdf for the forecast. In practice, we get a point estimate (the forecast) and a C.I.

• To get a point estimate, $\hat{Y}_{T+\ell}$, we need a cost function to judge various alternatives. This cost function is call *loss function*. Since we are working with forecast, we work with a expected loss function.

• A popular loss functions is the **Mean squared error** (**MSE**), which is quadratic and symmetric. We can use asymmetric functions, for example, functions that penalize positive errors more than negative errors.

For private use only - Not to be shared/posted online without authorization.

ARIMA: Forecasting – Optimal Forecast

• We derive the optimal forecast by minimizing the mean squared error (MSE):

$$MSE(e_{T+\ell}) = E[Y_{T+\ell} - \hat{Y}_{T+\ell}]^2$$

f.o.c.:

$$\frac{\delta E[Y_{T+\ell} - \hat{Y}_{T+\ell} \,|\, I_T]^2}{\delta\, \hat{Y}_{T+\ell}} = E[-2Y_{T+\ell} + 2\hat{Y}_{T+\ell} \,|\, I_T] = 0$$

$\Rightarrow$ Optimal forecast: $E[Y_{T+\ell} \,|\, I_T] = \hat{Y}_{T+\ell}$

• Different loss functions lead to different optimal forecast. For example, for the MAE, the optimal point forecast is the median.

• The computation of $E[Y_{T+\ell} \,|I_T]$ depends on the distribution of $\{\varepsilon_t\}$. Then, if

$$\{\varepsilon_t\} \sim \text{WN} \qquad \Rightarrow E[\varepsilon_{T+\ell} \,|I_T] = 0.$$

This assumption greatly simplifies computations.

Forecasting – Basic Concepts

• If

$$\{\varepsilon_t\} \sim \text{WN} \qquad \Rightarrow E[\varepsilon_{T+\ell} \,|I_T] = 0.$$

This assumption greatly simplifies computations, especially in the linear model.

• Then, for ARMA(p, q) stationary process (with a Wold representation), the minimum MSE linear forecast (best linear predictor) of $Y_{T+\ell}$, conditioning on I_T is:

$$Y_{T+\ell} = \theta_0 + \Psi_l \varepsilon_{T+\ell} + \Psi_{l+1} \varepsilon_{T+\ell-1} + \cdots$$

For private use only - Not to be shared/posted online without authorization.

Forecasting Steps for ARMA Models

- Data: $Y_1, Y_2, Y_3, ..., Y_T$

- Process:

(1) Find ARIMA model
(Use ACF, PACF or Minic)

$$Y_t = \phi Y_{t-1} + \varepsilon_t$$
$$\Downarrow$$

(2) Estimation
(& Evaluation in-sample)

$$\hat{\phi} \text{ (Estimate of } \phi)$$
$$\Downarrow$$
$$\hat{Y}_t = \hat{\phi} Y_{t-1} \text{ (Prediction)}$$

(3) Diagnostic Testing
(Check residuals, $\hat{\varepsilon}_t$, are WN)

$$\Downarrow$$
ACF & LB testing of $\hat{\varepsilon}_t$

(4) Forecast
(& Evaluation out-of-sample)

$$\Downarrow$$
$$\hat{Y}_{T+\ell} = \hat{\phi} \hat{Y}_{T+\ell-1} \text{(Forecast)}$$

Forecasting From ARMA Models

Example:

(1) Using AIC, we determine an AR(2) model.

$$Y_T = \mu + \phi_1 Y_{T-1} + \phi_2 Y_{T-2} + \varepsilon_T$$

(2) We use OLS to estimate μ, $\phi_1$and ϕ_2: $\hat{\mu}$, $\hat{\phi}_1$ & $\hat{\phi}_2$.

(3) We find residuals are WN.

(4) Now, we forecast. The one-step ahead forecast at time T:

$$\hat{Y}_{T+1} = E[Y_{T+1} \,|\, I_T = \{Y_T, Y_{T-1}, \dots, Y_1\}] = \hat{\mu} + \hat{\phi}_1 Y_T + \hat{\phi}_2 Y_{T-1}$$

At time $T+1$, we compute the one-step ahead forecast error, $e_T(1)$:

$$e_T(1) = Y_{T+1} - \hat{Y}_{T+1}$$

Note: After Q periods, we compute Q one-step ahead forecast errors and MSE.

8

For private use only - Not to be shared/posted online without authorization.

Forecasting From ARMA Models

• We observe the time series $I_T = \{Y_T, Y_{T-1}, \ldots, Y_1\}$.
- At time T, we want to forecast: Y_{T+1}, Y_{T+2}, …,$Y_{T+\ell}$.
- T: The forecast origin.
- ℓ: Forecast horizon
- $\hat{Y}_T(\ell)$: ℓ-*step ahead* forecast = Forecasted value Y_{T+l}

• Use the conditional expectation of $Y_{T+\ell}$, given the observed sample.

$$\hat{Y}_{T+\ell} = E[Y_{T+l} \,|\, Y_T, Y_{T-1}, \ldots, Y_1]$$

Example: One-step ahead forecast: $\hat{Y}_{T+1} = E[Y_{T+1} \,|\, Y_T, Y_{T-1}, \ldots, Y_1]$

• Forecast accuracy to be measured by MSE

$\Rightarrow$ conditional expectation, best forecast.

9

Forecasting From ARMA Models

• An ARMA forecasting is a combination of past $\hat{Y}_{T+\ell-i}$ forecasts and observed past $\hat{\varepsilon}_{t+\ell-i}$.

Example: We fit an ARMA(1, 2) model Y_t:

$$Y_t = \mu + \phi_1 Y_{t-1} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2}$$

• We want to produce at time T the forecast $Y_{T+\ell}$:

$$Y_{T+\ell} = \mu + \phi_1 Y_{T+\ell-1} + \varepsilon_{T+\ell} + \theta_1 \varepsilon_{T+\ell-1} + \theta_2 \varepsilon_{T+\ell-2}$$

• Two-step ahead forecast ($\ell = 2$): Conditional expectation.

$$\hat{Y}_{T+2} = \mu + \phi_1 E[Y_{T+1}|I_T] + E[\varepsilon_{T+2} \,|\, I_T] + \theta_1 \, \mathrm{E}[\varepsilon_{T+1} \,|\, I_T] + \theta_2 \mathrm{E}[\varepsilon_T \,|\, I_T]$$
$$= \mu + \phi_1 \hat{Y}_{t+1} + \theta_2 \, \hat{\varepsilon}_T$$

Actual: $Y_{T+2} = \mu + \phi_1 Y_{T+1} + \varepsilon_{T+2} + \theta_1 \varepsilon_{T+1} + \theta_2 \, \hat{\varepsilon}_T$

$e_T(2) = Y_{T+2} - \hat{Y}_{T+2} = \phi_1(\hat{Y}_{t+1} - Y_{T+1}) + \varepsilon_{T+2} + \theta_1 \varepsilon_{T+1}$

For private use only - Not to be shared/posted online without authorization.

Forecasting From ARMA Models

- We use the pure MA (**Wold**) representation of an ARMA(p, q):

$$\phi(L)(y_t - \mu) = \theta(L)\varepsilon_t$$

which involves inverting $\phi(L)$. That is,

$$(y_t - \mu) = \Psi(L)\varepsilon_t \Rightarrow \Psi(L) = \phi_p(L)^{-1}\theta_q(L)$$

- Then, the Wold representation:

$$Y_{T+\ell} = \mu + \varepsilon_{T+\ell} + \Psi_1\varepsilon_{T+\ell-1} + \Psi_2\varepsilon_{T+\ell-2} + \cdots + \Psi_\ell\ \varepsilon_T + \cdots$$

- The Wold representation depends on an infinite number of parameters, but, in practice, they decay rapidly.

- The forecast error is:

$$e_T(\ell) = \sum_{i=0}^{\ell-1} \Psi_i\ \varepsilon_{T+\ell-i} \qquad (\Psi_0 = 1)$$

Note: If E[$e_T(\ell)$] = 0, we say the forecast is **unbiased**.

Forecasting From ARMA Models

- The forecast error is:

$$e_T(\ell) = \sum_{i=0}^{\ell-1} \Psi_i\ \varepsilon_{T+\ell-i} \qquad (\Psi_0 = 1)$$

- The variance of the forecast error:

$$Var\big(e_T(\ell)\big) = Var\big(\textstyle\sum_{i=0}^{\ell-1} \Psi_i\varepsilon_{T+\ell-i}\big) = \sigma^2 \sum_{i=0}^{\ell-1} \Psi_i^2 \qquad (\Psi_0 = 1)$$

Example: One-step ahead forecast (ℓ = 1).

$$Y_{T+1} = \mu + \varepsilon_{T+1} + \Psi_1\varepsilon_T + \Psi_2\varepsilon_{T-1} + \Psi_3\varepsilon_{T-2} + \cdots$$

Forecast: $\hat{Y}_{T+1} = \mu + \Psi_1\varepsilon_T + \Psi_2\varepsilon_{T-1} + \cdots$

Forecast error: $e_T(1) = Y_{T+1} - \hat{Y}_{T+1} = \varepsilon_{T+1}$

Variance: $Var\big(e_T(1)\big) = \sigma^2$

For the two-step ahead forecast (ℓ = 2).

$$e_T(2) = Y_{T+2} - \hat{Y}_{T+2} = \varepsilon_{T+2} + \Psi_1\varepsilon_{T+1}$$

$$Var\big(e_T(2)\big) = \sigma^2 * (1 + \Psi_1^2)$$

For private use only - Not to be shared/posted online without authorization.

Forecasting From ARMA Models

• In the Wold representation, in practice, the parameters, Ψ_i's, decay rapidly. Then, as we forecast into the future, the forecasts tend to the unconditional forecasts, μ and σ^2:

$$\lim_{\ell \to \infty} \hat{Y}_T(\ell) = \mu$$

Not very interesting.

• This is why ARIMA forecasting is useful only for short-term.

Forecasting From ARMA Models: C.I.

• A 100(1 - α)% prediction interval for $Y_{T+\ell}$ (ℓ-steps ahead) is

$$\hat{Y}_T(\ell) \pm z_{\alpha/2} \sqrt{Var(e_T(\ell))}$$

or, $$\hat{Y}_T(\ell) \pm z_{\alpha/2}\, \sigma \sqrt{\textstyle\sum_{i=0}^{\ell-1} \Psi_i^2}$$

Example: 95% C.I. for the 2-step-ahead forecast:

$$\hat{Y}_T(2) \pm 1.96\, \sigma \sqrt{1 + \Psi_1^2}$$

• When computing prediction intervals from data, we substitute estimates for parameters, giving approximate prediction intervals.

Note: MSE[$\varepsilon_{T+\ell}$] = MSE[$e_{T+\ell}$] = $\sigma^2 \sum_{i=0}^{\ell-1} \Psi_i^2$

For private use only - Not to be shared/posted online without authorization.

Forecasting From ARMA Model: Updating

• Suppose we have T observations at time $t = T$. We have a good ARMA model for Y_T. We obtain the forecast for Y_{T+1}, Y_{T+2}, etc.

• At $t = T + 1$, we observe Y_{T+1}. Now, we update our forecasts using the original value of Y_{T+1} and the forecasted value of it.

• The forecast error is:

$$e_T(\ell) = Y_{T+\ell} - \hat{Y}_T(\ell) = \sum_{i=0}^{\ell-1} \Psi_i \, \varepsilon_{T+\ell-i}$$

• We can also write this as

$$\begin{aligned} e_{T-1}(\ell+1) &= Y_{T-1+\ell+1} - \hat{Y}_{T-1}(\ell+1) \\ &= \sum_{i=0}^{\ell} \Psi_i \, \varepsilon_{T-1+\ell+1-i} \\ &= \sum_{i=0}^{\ell} \Psi_i \, \varepsilon_{T+\ell-i} \\ &= \sum_{i=0}^{\ell-1} \Psi_i \, \varepsilon_{T+\ell-i} + \Psi_\ell \, \varepsilon_T \\ &= e_T(\ell) + \Psi_\ell \, \varepsilon_T \end{aligned}$$

Forecasting From ARMA Model: Updating

• Then,

$$\begin{aligned} Y_{T+\ell} - \hat{Y}_{T-1}(\ell+1) &= Y_{T+\ell} - \hat{Y}_T(\ell) + \Psi_\ell \, \varepsilon_T \\ \hat{Y}_T(\ell) &= \hat{Y}_{T-1}(\ell+1) + \Psi_\ell \, \varepsilon_T \\ &= \hat{Y}_{T-1}(\ell+1) + \Psi_\ell\{Y_T - \hat{Y}_{T-1}(1)\} \\ \Rightarrow \hat{Y}_{T+1}(\ell) &= \hat{Y}_T(\ell+1) + \Psi_\ell\{Y_{T+1} - \hat{Y}_T(1)\} \end{aligned}$$

Example: $\ell = 1, T = 100.$

$$\hat{Y}_{101}(1) = \hat{Y}_{100}(2) + \Psi_1\{Y_{101} - \hat{Y}_{100}(1)\}$$

For private use only - Not to be shared/posted online without authorization.

Forecasting From ARMA Model: Transformations

• If we use variance stabilizing transformation, after the forecasting, we need to convert the forecasts for the original series.

• For example, if we use log-transformation, then,

$$E[Y_{T+\ell} |\ I_T] \geq \exp\{E[\ln(Y_{T+\ell})|\ I_T]\}$$

• If $X \sim N(\mu, \sigma^2)$, then, $E[\exp(X)] = e^{\mu + \frac{\sigma^2}{2}}$

• The MSE forecast for the original series is:

$$\exp\left[\underbrace{\hat{Z}_n(\ell)}_{\mu = E(Z_{n+\ell}|Z_1,\cdots,Z_n)} + \frac{1}{2}\underbrace{Var(e_n(\ell))}_{\sigma^2 = Var(Z_{n+\ell}|Z_1,\cdots,Z_n)}\right] \quad \text{where } Z_{n+\ell} = \ln(Y_{n+\ell})$$

Forecasting From ARMA Model: Remarks

• In general, we need a large T. Better estimates and it is possible to check for model stability and check forecasting ability of model by withholding data.

• Seasonal patterns also need large T. Usually, you need 4 to 5 seasons to get reasonable estimates.

• Parsimonious models are very important. Easier to compute and interpret models and forecasts. Forecasts are less sensitive to deviations between parameters and estimates.

For private use only - Not to be shared/posted online without authorization.

Forecasting From Simple Models: ES

• Industrial companies, with a lot of inputs and outputs, want quick and inexpensive forecasts. Easy to fully automate. In general, we use past Y_t to forecast future Y_t's, usually referred as the **level's forecasts**.

• Exponential Smoothing Models (ES) fulfill these requirements.

• In general, these models are limited and not optimal, especially compared with Box-Jenkins methods.

• Goal of these models: Suppress the short-run fluctuation by smoothing the series. For this purpose, a weighted average of all previous values works well.

• There are many ES models. We will go over the Simple Exponential Smoothing (**SES**) & Holt-Winter's Exponential Smoothing (**HW ES**).

SES: Forecast and Updating

• From the updating equation S_t:

$$S_t = S_{t-1} + \alpha\,(Y_{t-1} - S_{t-1})$$

we compute the forecast for next period $(t+1)$:

$$S_{t+1} = S_t + \alpha(Y_t - S_t) \qquad (\hat{Y}_{t+1} = S_{t+1})$$

That is, a simple updating forecast: last period forecast + adjustment.

• The forecast for the period $t+2$, we have:

$$S_{t+2} = S_{t+1} + \alpha(Y_{t+1} - S_{t+1}) = S_{t+1}$$

• The ℓ-step ahead forecast is:

$$S_{t+\ell} = S_{t+1} \qquad \Rightarrow \text{A naive forecast!}$$

Note: SES forecasts are not very interesting after $\ell > 1$.

For private use only - Not to be shared/posted online without authorization.

SES: Exponential?

- Q: Why Exponential?

For the observed time series $\{Y_1, Y_2, \ldots, Y_t, Y_{t+1}\}$, using backward substitution, $S_{t+1} = \hat{Y}_t(1)$ can be expressed as a weighted sum of previous observations:

$$S_{t+1} = \alpha Y_t + (1-\alpha) S_t = \alpha Y_t + (1-\alpha)[\alpha Y_{t-1} + (1-\alpha) S_{t-1}]$$
$$= \alpha Y_t + \alpha(1-\alpha) Y_{t-1} + (1-\alpha)^2 S_{t-1}$$
$$\Rightarrow \hat{Y}_t(1) = S_{t+1} = c_0 Y_t + c_1 Y_{t-1} + c_2 Y_{t-2} + \cdots$$

where c_i's are the weights, with

$$c_i = \alpha(1-\alpha)^i; \; i = 0, 1, \quad \ldots; \; 0 \leq \alpha \leq 1.$$

- We have decreasing weights, by a constant ratio for every unit increase in lag.

21

SES: Forecast and Updating

Example: An industrial firm uses SES to forecast sales:

$$S_{t+1} = S_t + \alpha * (Y_t - S_t)$$

The firm estimates α = **0.25**. The firm observes Y_t = **5** and, last period's forecast, S_t = **3**.

Then, the forecast for time $t+1$ is:

$$S_{t+1} = \mathbf{3} + \mathbf{0.25} * (\mathbf{5} - \mathbf{3}) = \mathbf{3.50}$$

The forecast for time $t+1$ (& any period after time $t+1$) is:

$$S_{t+\ell} = S_{t+1} = \mathbf{3.50} \qquad \text{for } \ell > 1.$$

Later, the firm observes: Y_{t+1} = **4.77**, Y_{t+2} = **3.15**, & Y_{t+3} = **1.85**. Then, the MSE:

MSE = $\frac{1}{3}$ * [(**4.77** – **3.50**)^2 + (**3.15** - **3.50**)^2 + (**1.85** - **3.50**)^2] = **1.486**.

For private use only - Not to be shared/posted online without authorization.

SES: Forecast and Updating

Example (continuation):

Note: If α = 0.75, then

$$S_{t+1} = 3 + 0.75 * (5 – 3) = 4.50$$

A bigger α gives more weight to the more recent observation –i.e., Y_t.

Again, the forecast for time $t+1$ (& any period after time $t+1$) is:

$$S_{t+\ell} = S_{t+1} = 4.50 \qquad \text{for } \ell > 1.$$

SES: Selecting α

- Choose α between 0 and 1.

- If α = 1, it becomes a naive model; if α ≈ 1, more weights are put on recent values. The model fully utilizes forecast errors.

- If α is close to 0, distant values are given weights comparable to recent values. Set α ≈ 0 when there are big random variations in Y_t.

- α is often selected as to minimize the MSE.

- In empirical work, 0.05 ≤ α ≤ 0.3 are used (α ≈ 1 is used rarely).

Numerical Minimization Process:

- Take different α values ranging between 0 and 1.

- Calculate 1-step-ahead forecast errors for each α.

- Calculate MSE for each case.

Choose α which has the min MSE: $e_t = Y_t - S_t \Rightarrow \min \sum_{t=1}^{n} e_t^2 \Rightarrow \alpha$

24

For private use only - Not to be shared/posted online without authorization.

SES: Selecting α – MSE

$$S_{t+1} = \alpha Y_t + (1-\alpha) S_t$$

Time	Y_t	S_{t+1} (α=0.10)	$(Y_t - S_t)^2$
1	5	-	-
2	7	(0.1)5 +(0.9)5 = 5	4
3	6	(0.1)7 + (0.9)5 = 5.2	0.64
4	3	(0.1)6 + (0.9)5.2 = 5.28	5.1984
5	4	(0.1)3 + (0.9)5.28 = 5.052	1.107
		TOTAL	**10.945**

$$MSE = \frac{SSE}{n-1} = 2.74$$

• Calculate this for α = 0.2, 0.3,…, 0.9, 1 and compare the MSEs. Choose α with minimum MSE.

Note: $Y_{t-1} = 5$ is set as the initial value for the recursive equation.

25

SES: Initial Values

• We have a recursive equation, we need initial values, S_1 (or Y_0).

• Approaches:

– Set S_1 equal to Y_1. Then, $S_2 = Y_1$.

– Take the average of, say first 4 or 5 observations. Then, we start forecasting at time 5 or 6, respectively.

– Estimate S_1 (similar to the estimation of α.)

26

For private use only - Not to be shared/posted online without authorization.

SES: Forecasting Examples

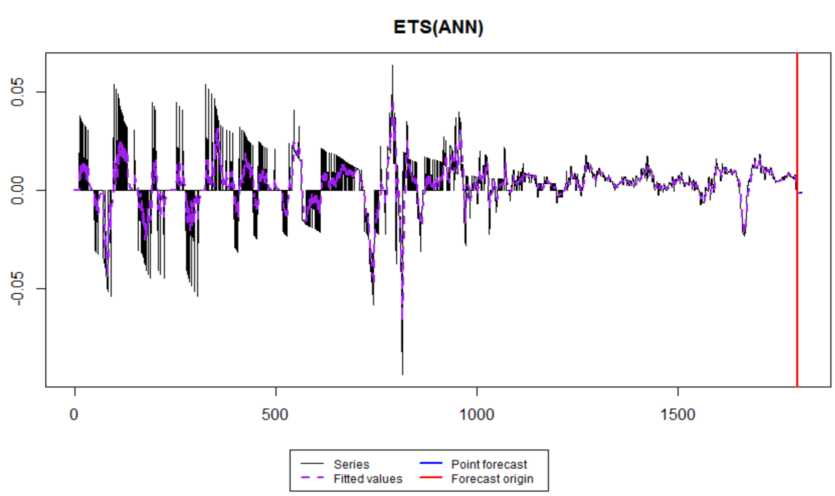
Example 1: We want to forecast log changes in **U.S. monthly dividends** (T=1796) using SES. First, we estimate the model using the R function *HoltWinters*(), which has as a special case SES: set beta=FALSE, gamma=FALSE. We use estimation period T=1750.

```
mod1 <- HoltWinters(lr_d[1:1750], beta=FALSE, gamma=FALSE)
> mod1
Holt-Winters exponential smoothing without trend and without seasonal component.

Call:
HoltWinters(x = lr_d[1:1750], beta = FALSE, gamma = FALSE)

Smoothing parameters:
 alpha: 0.289268                          ⇒ Estimated α
 beta : FALSE
 gamma: FALSE

Coefficients:
        [,1]
a 0.004666795                             ⇒ Forecast
```

27

SES: Forecasting Examples

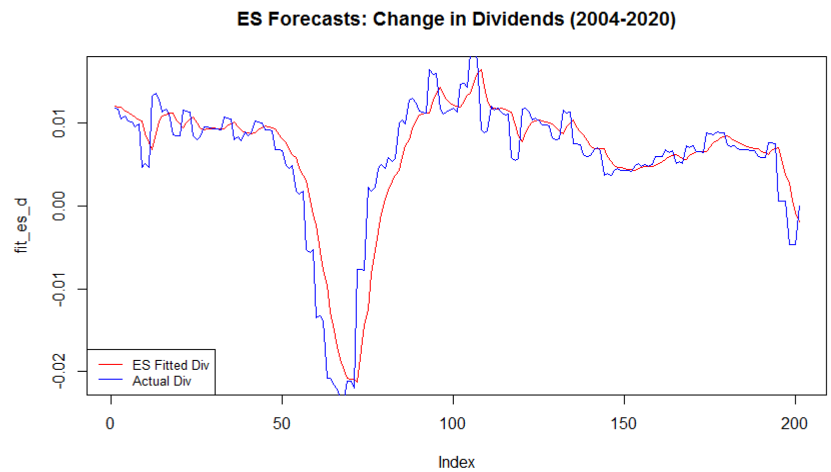
Example 1 (continuation):



28

SES: Forecasting Examples

Example 1 (continuation):



29

SES: Forecasting Examples

Example 1 (continuation): Now, we do one-step ahead forecasts

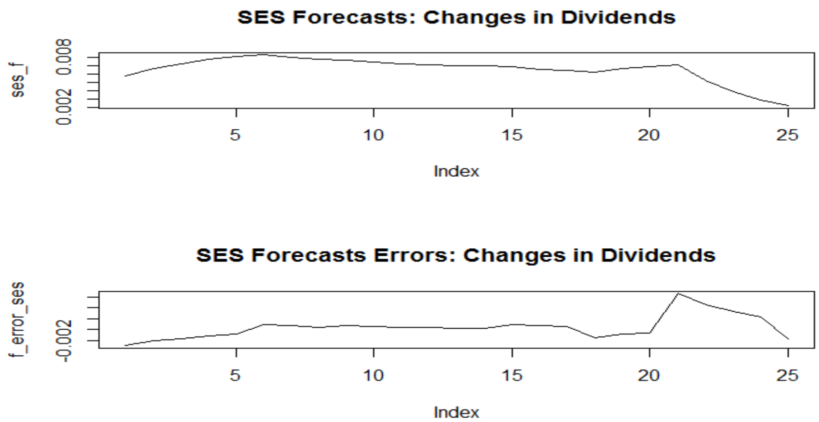
```
T_last <- nrow(mod1$fitted)                    # number of in-sample forecasts
h <- 25                                        # forecast horizon
ses_f <- matrix(0,h,1)                         # Vector to collect forecasts
alpha <- 0.29
y <- lr_d
T <- length(lr_d)
sm <- matrix(0,T,1)
T1 <- T – h + 1                                # Start of forecasts
a <- T1                                        # index for while loop
sm[a-1] <- mod1$fitted[T_last]                 # last in-sample forecast
while (a <= T) {
        sm[a] = alpha * y[a-1] + (1-alpha) * sm[a-1]
a <- a + 1
}

ses_f <- sm[T1:T]
ses_f
f_error_ses <- sm[T1:T] - y[T1:T]              # forecast errors
MSE_ses <- sum(f_error_ses^2)/h                # MSE
plot(ses_f, type="l", main ="SES Forecasts: Changes in Dividends")
```

30

SES: Forecasting Examples

Example 1 (continuation):

```
> ses_f
f_error_ses <- sm[T1:T] - y[T1:T]
> plot(ses_f, type="l", main ="SES Forecasts: Changes in Dividends")
```



SES: Forecasting U.S. Dividends

Example 1 (continuation): *h-step-ahead* forecasts

```
> forecast(mod1, h=25, level=.95)
     Point Forecast       Lo 95      Hi 95
1751    0.004666795 -0.01739204 0.02672563
1752    0.004666795 -0.01829640 0.02762999
1753    0.004666795 -0.01916647 0.02850006
1754    0.004666795 -0.02000587 0.02933947
1755    0.004666795 -0.02081765 0.03015124
1756    0.004666795 -0.02160435 0.03093794
1757    0.004666795 -0.02236816 0.03170175
1758    0.004666795 -0.02311098 0.03244457
1759    0.004666795 -0.02383445 0.03316804
1760    0.004666795 -0.02454001 0.03387360
1761    0.004666795 -0.02522891 0.03456250
1762    0.004666795 -0.02590230 0.03523589
1763    0.004666795 -0.02656117 0.03589476
1764    0.004666795 -0.02720642 0.03654001
...
```

Note: Constant forecasts, but C.I. gets wider (as expected) with h. [32]

SES: Forecasting Examples

Example 2: We want to forecast **log monthly U.S. vehicles** (1976-2020, T=537) using SES.

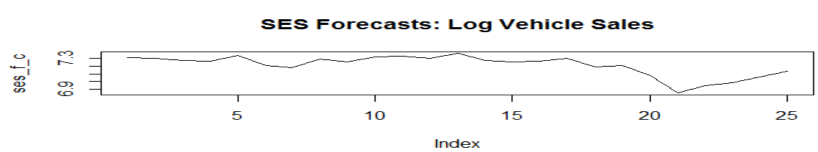
```
mod_car <- HoltWinters(l_car[1:512], beta=FALSE, gamma=FALSE)
> mod_car
Holt-Winters exponential smoothing without trend and without seasonal component.

Call:
HoltWinters(x = l_car[1:512], beta = FALSE, gamma = FALSE)

Smoothing parameters:
 alpha: 0.4888382                          ⇒ Estimated α
 beta : FALSE
 gamma: FALSE

Coefficients:
      [,1]
a 7.315328
```

33

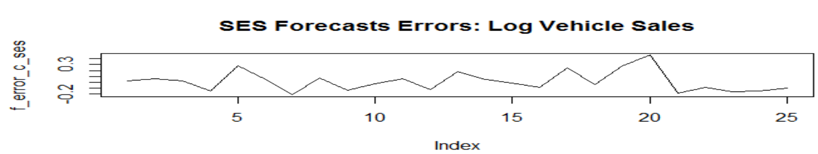
SES: Forecasting Examples

Example 2 (continuation): Now, we do one-step ahead forecasting

```
ses_f_c <- sm_c[T1:T]
f_error_c_ses <- sm_c[T1:T] - y[T1:T]
> plot(ses_f_c, type="l", main ="SES Forecasts: Log Vehicle Sales")
```




```
> plot(f_error_c_ses, type="l", main ="SES Forecasts Errors: Log Vehicle Sales")
```




```
MSE_ses <- sum(f_error_c_ses^2)/h
> MSE_ses
[1] 0.027889
```

34

For private use only - Not to be shared/posted online without authorization.

SES: Remarks

- Some computer programs automatically select the optimal α, using a line search method or non-linear optimization techniques (R does this with function *HoltWinters*).

- We have a recursive equation, we need initial values for S_1. Using an average of the first observations is common.

- This model ignores trends or seasonalities. Not very realistic, especially for manufacturing facilities, retail sector, and warehouses.

- Deterministic components, D_t, can be easily incorporated.

- The model that incorporates both a trend and seasonal features is called *Holt-Winter's ES*.

35

Holt-Winters (HW) Exponential Smoothing

- In the model for Y_t, in addition to the level (S_t), we introduce **trend** (T_t) & **seasonality** (I_t) factors. Since we produce smooth forecasts for T_t & I_t, this method is also called *triple exponential smoothing*.

- The h-step ahead forecast is a combination of the smooth forecasts of S_t (**Level**), T_t (**Trend**) & I_{t+h-s} (**Seasonal**).

- Both, T_t & I_t, can be included as *additively* or *multiplicatively* factors. In this class, we consider an additive trend and the seasonal factor as additive or multiplicative. We produce *h*-step ahead forecasts:

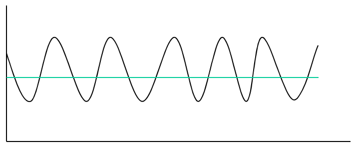
- For the additive model: $\hat{Y}_t(h) = S_t + h\,T_t + I_{t+h-s}$
- For the multiplicative model: $\hat{Y}_t(h) = (S_t + h\,T_t) * I_{t+h-s}$

Note: Seasonal factor is multiplied in the h-step ahead forecast.

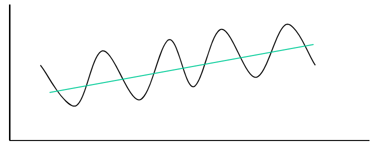
36

For private use only - Not to be shared/posted online without authorization.

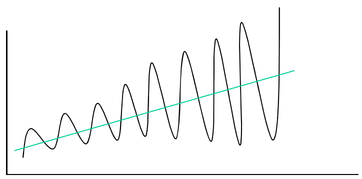
Holt-Winters (HW) ES: Trend & Seasonality



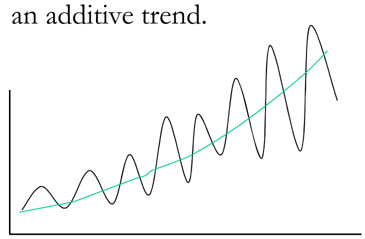
1. No trend and additive seasonal variability.



2. Additive seasonal variability with an additive trend.



3. Multiplicative seasonal variability with an additive trend.



4. Multiplicative seasonal variability with a multiplicative trend.

<u>Note</u>: We will use Model 2 (Additive) and Model 3 (Multiplicative).

Holt-Winters (HW) ES: Additive

• Additive model (additive trend & additive seasonality) forecast:

$$\hat{Y}_t(h) = S_t + h\,T_t + I_{t+h-s}$$

where s is the number of periods in seasonal cycles (=4 for quarters).

• Components:

- **The level**, S_t: A weighted average of "*seasonal adjusted*" Y_t (=$Y_t - I_{t-s}$), and the non-seasonal forecast $(S_{t-1} + T_{t-1})$:

$$S_t = \alpha(Y_t - I_{t-s}) + (1-\alpha)(S_{t-1} + T_{t-1})$$

- **The trend**, T_t: A weighted average of T_{t-1} and the change in S_t.

$$T_t = \beta(S_t - S_{t-1}) + (1-\beta)T_{t-1}$$

- **The seasonality**, I_t: A weighted average of seasonal index of s last year, I_{t-s}, and the current seasonal index $(Y_{t-1} - S_{t-1} - T_{t-1})$:

$$I_t = \gamma(Y_t - S_{t-1} - T_{t-1}) + (1-\gamma)I_{t-s}$$

38

For private use only - Not to be shared/posted online without authorization.

Holt-Winters (HW) ES: Additive

• Then, the model for the h-step ahead forecast

$$\hat{Y}_t(h) = S_t + h\,T_t + I_{t+h-s}$$

has three equations:

Level: $S_t = \alpha(Y_t - I_{t-s}) + (1-\alpha)(S_{t-1} + T_{t-1})$

Trend: $T_t = \beta\,(S_t - S_{t-1}) + (1-\beta)\,T_{t-1}$

Seasonal: $I_t = \gamma(Y_t - S_{t-1} - T_{t-1}) + (1-\gamma)I_{t-s}$

• We have only three smoothing parameters:

α = level coefficient

β = trend coefficient

γ = seasonality coefficient

39

Holt-Winters (HW) ES: Multiplicative

• In the multiplicative seasonal case (with an additive trend), we have the h-step ahead forecast:

$$\hat{Y}_t(h) = (S_t + h\,T_t) * I_{t+h-s}$$

• Details for *multiplicative* seasonality –i.e., Y_t/I_t– and *additive* trend

- The forecast, S_t, now shows the average Y_t adjusted ($\frac{Y_t}{I_{t-s}}$).
- The trend, T_t, is a weighted average of T_{t-1} and the change in S_t.
- The seasonality is also a weighted average of I_{t-s} and the Y_t/S_t.

• Then, the model has three equations:

$$S_t = \alpha\,\frac{Y_t}{I_{t-s}} + (1-\alpha)\,(S_{t-1} + T_{t-1})$$
$$T_t = \beta\,(S_t - S_{t-1}) + (1-\beta)\,T_{t-1}$$
$$I_t = \gamma\,\frac{Y_t}{S_t} + (1-\gamma)\,I_{t-s}$$

40

For private use only - Not to be shared/posted online without authorization.

Holt-Winters (HW) ES: Multiplicative

- We think of (Y_t/S_t) as capturing *seasonal effects*.
  s = # of periods in the seasonal cycles
  (s = 4, for quarterly data; s = 12, for monthly)

- Again, we have only three parameters:
  α = smoothing parameter
  β = trend coefficient
  γ = seasonality coefficient

- Q: How do we determine these 3 parameters?

- Ad-hoc method: α, β and γ can be chosen as values between
  $0.02 < \alpha, \gamma, \beta < 0.2$
- Optimal method: Minimization of the MSE, as in SES.

41

Holt-Winters (HW) ES: Multiplicative

Example: An industrial firm uses HW ES to forecast sales next two quarters (h = 1, 2, & 3; with s = 4):

$$\hat{Y}_t(h) = \hat{Y}_{t+h} = (S_t + h\,T_t) * I_{t+h-s}$$

with S_t, T_t, & I_t factors given by:

$$S_t = \alpha \frac{Y_t}{I_{t-s}} + (1-\alpha)(S_{t-1} + T_{t-1})$$
$$T_t = \beta\,(S_t - S_{t-1}) + (1-\beta)\,T_{t-1}$$
$$I_t = \gamma \frac{Y_t}{S_t} + (1-\gamma)\,I_{t-s}$$

The firm estimates: α = 0.25; β = 0.1; & γ = 0.4. It observes Y_t = 5; last quarter's smoothed forecasts: S_{t-1} = 3, T_{t-1} = 1.2; & last year's seasonal factors: I_{t-4} = 1.1, I_{t-3} = 0.7, I_{t-2} = 1.2, & I_{t-1} = 0.8.

- Components forecasts:

$$S_t = 0.25\,\frac{5}{1.1} + (1-0.25) * (3+1.3) = 4.2864$$

For private use only - Not to be shared/posted online without authorization.

Holt-Winters (HW) ES: Multiplicative

Example (continuation): (α = 0.25; β = 0.1; & γ = 0.4.)

$$S_t = 0.25 * \frac{5}{1.1} + (1 - 0.25) * (3 + 1.2) = 4.2864$$

$$T_t = 0.1 * (4.2864 - 3) + (1 - 0.1) * 1.2 = 1.2086$$

$$I_t = 0.4 * \frac{5}{4.2864} + (1 - 0.4) * 1.1 = 1.1266$$

The forecast for h = 1 (next quarter) is:

$$\hat{Y}_{t+1} = (4.2864 + 1.2086) * 0.7 = 4.8125$$

The forecast for h = 2 & 3 are:

$$\hat{Y}_{t+2} = (4.2864 + 2 * 1.2086) * 1.2 = 7.8475.$$

$$\hat{Y}_{t+3} = (4.2864 + 3 * 1.2086) * 0.8 = 6.1329.$$

HW ES: Initial Values

- Initial values for algorithm

- We need at least one complete season of data to determine the initial estimates of I_{t-s}.

- Initial values for *multiplicative* model:

$$S_0 = \sum_{t=1}^{s} Y_t / s$$

$$T_0 = \frac{1}{s}\left(\frac{Y_{s+1} - Y_1}{s} + \frac{Y_{s+2} - Y_2}{s} + \cdots + \frac{Y_{s+s} - Y_s}{s}\right)$$

$$\text{or } T_0 = \left[\left\{\sum_{t=1}^{s} Y_t / s\right\} - \left\{\sum_{t=s+1}^{2s} Y_t / s\right\}\right] / s$$

44

For private use only - Not to be shared/posted online without authorization.

HW ES: Initial Values

• Algorithm to compute initial values for seasonal component I_s. Assume we have T observation and quarterly seasonality (s=4):

(1) Compute the averages of each of T years.

$$A_t = \sum_{i=1}^{4} Y_{t,i}/4\,, \quad t = 1, 2, \cdots, 6 \quad \text{(yearly averages)}$$

(2) Divide the observations by the appropriate yearly mean: $Y_{t,i}/A_t$.

(3) I_s is formed by computing the average $Y_{t,i}/A_t$ per year:

$$I_s = \sum_{i=1}^{T} Y_{t,s}/A_t \quad s = 1, 2, 3, 4$$

45

HW ES: Damped Model

• We can damp the trend as the forecast horizon increases, using a parameter ϕ. For the multiplicative model we have:

$$S_t = \alpha \frac{Y_t}{I_{t-s}} + (1-\alpha)(S_{t-1} - \phi\, T_{t-1})$$
$$T_t = \beta(S_t - S_{t-1}) + (1-\beta)T_{t-1}$$
$$I_t = \gamma \frac{Y_t}{S_t} + (1-\gamma)I_{t-s}$$

• *h-step ahead* forecast:

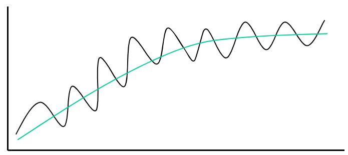
$$\hat{Y}_t(h) = \{S_t + \left(1 + \phi + \phi^2 + \cdots + \phi^{2h-1}\right)T_t\} * I_{t+h-s}$$

• This model is based on practice: It seems to work well for industrial outputs. Not a lot of theory or clear justification behind the damped trend.

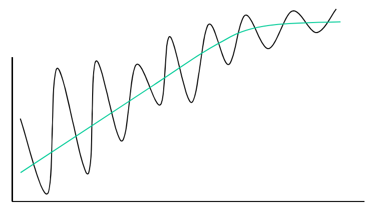
46

For private use only - Not to be shared/posted online without authorization.

ES Models: Damped Model – Types



5. Dampened trend with additive seasonal variability.



6. Multiplicative seasonal variability and dampened trend.

• Overall, we have different models, incorporating different features:
- Trend: Additive or multiplicative, dampened or not
- Seasonal variability: Additive or multiplicative

• Q: With all these models, which one we should use? It depends on the data at hand.

HW ES: Example – Log U.S. Vehicles Sales

Example: We want to forecast log U.S. monthly vehicle sales with HW. We use the R function *HoltWinters*().

```
l_car_18 <- l_car[1:512]
l_car_ts <- ts(l_car_18, start = c(1976, 1), frequency = 12)     # convert lr_d in a ts object
hw_d_car <- HoltWinters(l_car_18, seasonal="additive")
> hw_d_car
Holt-Winters exponential smoothing with trend and additive seasonal component.

Call:
HoltWinters(x = lr_d_ts, seasonal = "additive")

Smoothing parameters:
 alpha: 0.4355244                  ⇒ Estimated smoothing parameter
 beta : 0.009373815                ⇒ Estimated trend parameter ≈ 0 (no trend)
 gamma:0.3446495                   ⇒ Estimated seasonal parameter
```

48

For private use only - Not to be shared/posted online without authorization.

HW ES: Example – Log U.S. Vehicles Sales

Example (continuation):

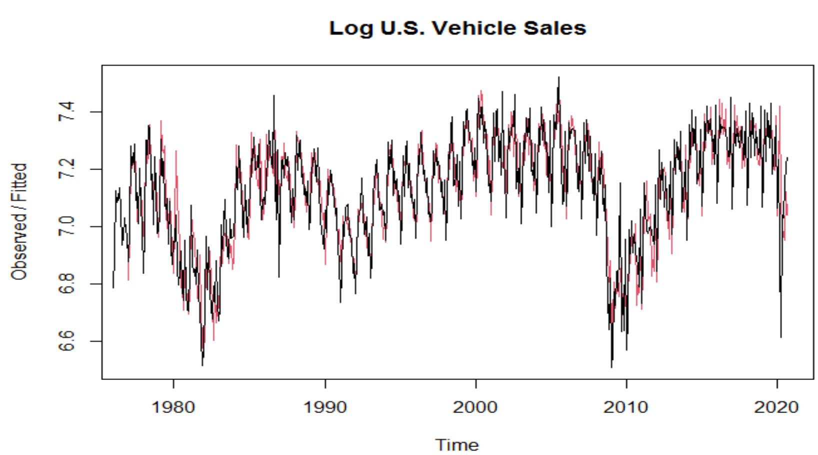
```
> hw_d_car

Coefficients:
            [,1]
a    7.177857555      ⇒ forecast for level
b    0.0001100345     ⇒ forecast for trend
s1  -0.075314457      ⇒ forecast for seasonal month 1
s2  -0.084468361      ⇒ forecast for seasonal month 2
s3   0.049447067
s4  -0.273299309
s5  -0.138251757
s6  -0.026603921
s7  -0.144953062
s8   0.079214066
s9   0.037899454
s10  0.020477134
s11  0.089309775
s12 -0.012530316
```

49

HW ES: Example – Log U.S. Vehicles Sales

Example (continuation):

plot(hw_d_car)



For private use only - Not to be shared/posted online without authorization.

SES: Forecasting Log U.S. Vehicles Sales

Example (continuation): Now, we forecast one-step ahead forecasts

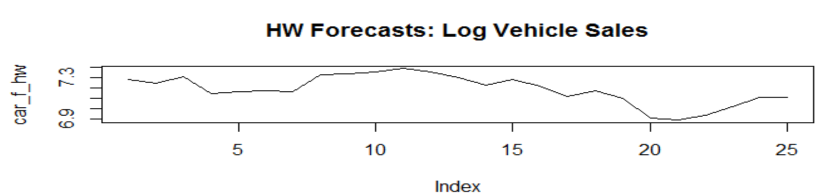
```
T_last <- nrow(hw_d_car$fitted)
h <- 25
ses_f_hw <- matrix(0,h,1)
 alpha <- 0.4355244
 beta <- 0.009373815
 gamma <- 0.3446495
y <- l_car
T <- length(l_car)
sm <- matrix(0,T,1)
Tr <- matrix(0,T,1)
I <- matrix(0,T,1)
T1 <- T-h+1
a <- T1
sm[a-1] <- 7.177857555
Tr[a-1] <- -0.000309358
I[501:512] <- c(-0.075314457,-0.084468361,0.049447067,-0.273299309,-0.138251757, -
0.026603921, -0.144953062,0.079214066,0.037899454,0.020477134,0.089309775,-
0.012530316)
```

51

SES: Forecasting Log U.S. Vehicles Sales

Example (continuation):

```
while (a <= T) {
        sm[a] = alpha * y[a-1] +  (1-alpha) * sm[a-1]
        Tr[a] = beta * (sm[a] - sm[a-1]) + (1 - beta) * Tr[a-1]
        I[a] = gamma * (y[a] - sm[a]) + (1 - gamma) * I[a - 12]
a <- a + 1
}
hh <- c(1:h)
car_f_hw <- sm[T1:T] + hh*Tr[T1:T] + I[T1:T]
car_f_hw
f_error_c_hw<- car_f_hw - y[T1:T]
plot(car_f_hw, type="l", main ="SES Forecasts: Log Vehicle Sales")
```

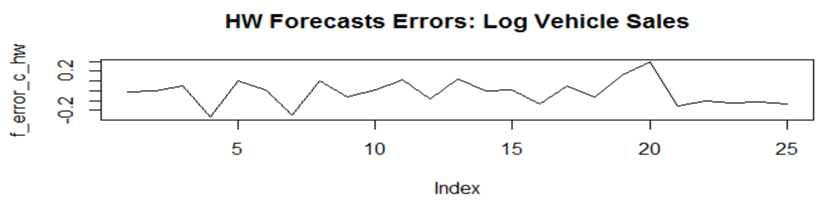


For private use only - Not to be shared/posted online without authorization.

SES: Forecasting Log U.S. Vehicles Sales

Example (continuation):

plot(f_error_c_hw, type="l", main ="SES Forecasts Errors: Log Vehicle Sales")



MSE_hw <- sum(f_error_c_hw^2)/h
> MSE_hw
[1] 0.01655964

HW ES: Remarks

- Remarks

- If a computer program selects $\gamma = 0 = \beta$, it has a lack of trend or seasonality. It implies a constant (deterministic) component. In this case, an ARIMA model with deterministic trend may be a more appropriate model.

- For HW ES, a seasonal weight near one implies that a non-seasonal model may be more appropriate.

- We can model seasonalities as multiplicative or additive:

$\Rightarrow$ Multiplicative seasonality: Forecast$_t$ = $S_t * I_{t-s}$.

$\Rightarrow$ Additive seasonality: Forecast$_t$ = $S_t + I_{t-s}$.

For private use only - Not to be shared/posted online without authorization.

Evaluation of forecasts: Accuracy measures

• The mean squared error (*MSE*) and mean absolute error (*MAE*) are the most popular accuracy measures:

$$\text{MSE} = \frac{1}{m}\sum_{i=T+1}^{T+m}(\hat{y}_i - y_i)^2 = \frac{1}{m}\sum_{i=T+1}^{T+m} e_i{}^2$$

$$\text{MAE} = \frac{1}{m}\sum_{i=T+1}^{T+m} |\hat{y}_i - y_i\,| = \frac{1}{m}\sum_{i=T+1}^{T+m} |e_i|$$

where m is the number of out-of-sample forecasts.

• But other measures are routinely used:

- Mean absolute percentage error (*MAPE*) = $\frac{100}{T-(m-1)}\sum_{i=T+1}^{T+m} |\frac{\hat{y}_i - y_i}{y_i}|$

- Absolute *MAPE (AMAPE)* = $\frac{100}{T-(m-1)}\sum_{i=T+1}^{T+m} |\frac{\hat{y}_i - y_i}{\hat{y}_i + y_i}|$

Remark: There is an asymmetry in MAPE, the level y_i matters.

Evaluation of forecasts: Accuracy measures

- % correct sign predictions (PCSP) = $\frac{1}{T-(m-1)}\sum_{i=T+1}^{T+m} z_i$

where z_i = 1 if $(\hat{y}_{i+l} * y_{i+l}) > 0$
= 0, otherwise.

- % correct direction change predictions (PCDP)= $\frac{1}{T-(m-1)}\sum_{i=T+1}^{T+m} z_i$

where z_i = 1 if $(\hat{y}_{i+l} - y_i) * (y_{i+l} - y_i) > 0$
= 0, otherwise.

Remark: We value forecasts with the right direction (sign) or forecast that can predict turning points. For stock investors, the sign matters!

• MSE penalizes large errors more heavily than small errors, the sign prediction criterion, like MAE, does not penalize large errors more.

For private use only - Not to be shared/posted online without authorization.

Evaluation of forecasts: Accuracy measures

Example: We compute MSE and the % of correct direction change (PCDC) predictions for the one-step forecasts for U.S. monthly vehicles sales based on the SES and HW ES models.

```
> MSE_ses
[1] 0.027889

> MSE_hw
[1] 0.01655964
```

• We calculate PCDC with following script for HW & SES:

```
bb_hw <- (car_f_hw - y[(T1-1):(T-1)]) * (y[T1:T] - y[(T1-1):(T-1)])
indicator_hw <- ifelse(bb_hw > 0,1,0)     # ifelse ("if else") produces a 1 if condition is true
pcdc_hw <- sum(indicator_hw)/h
> indicator_hw
 [1] 1 1 1 0 1 1 1 1 1 1 0 1 1 1 1 0 1 1 1 1 1 1 0 0 0
> pcdc_hw
[1] 0.76
```

Evaluation of forecasts: Accuracy measures

Example (continuation):

```
bb_s <- (ses_f_c - y[(T1-1):(T-1)]) * (y[T1:T] - y[(T1-1):(T-1)])
indicator_s <- ifelse(bb_s > 0,1,0)
pcdc_s <- sum(indicator_s)/h
> indicator_s
 [1] 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 0 1 1 0 0 0
> pcdc_s
[1] 0.76
```

Note: Same percentage of correct direction change (PCDC) predictions, but the sequence of correct predictions is not the same.

For private use only - Not to be shared/posted online without authorization.

Evaluation of forecasts: DM Test

- To determine if one model predicts better than another, we define the loss differential between two forecasts:

$$d_t = g(e_t^{M1}) - g(e_t^{M2})$$

where g(.) is the forecasting loss function, M1 and M2 are two competing sets of forecasts –could be from models or something else.

- We only need $\{e_t^{M1}\}$ & $\{e_t^{M2}\}$, not the structure of M1 or M2. In this sense, this approach is "*model-free*."

- Typical (symmetric) loss functions: $g(e_t) = e_t^2$ & $g(e_t) = |e_t|$.

- But other g(.)'s can be used: $g(e_t) = \exp(\lambda\, e_t^2) - \lambda\, e_t^2$ ($\lambda>0$).

Note: This is a more general test than MGN: It works for any loss function, not just MSE.

Evaluation of forecasts: DM Test

- Then, we test the null hypotheses of equal predictive accuracy:

$$H_0: E[d_t] = 0$$
$$H_1: E[d_t] = \mu \neq 0.$$

- Diebold and Mariano (1995) assume $\{e_t^{M1}\}$ & $\{e_t^{M2}\}$ is covariance stationarity and other regularity conditions (finite Var[d_t], independence of forecasts after ℓ periods) needed to apply CLT. Then,

$$\frac{\bar{d}-\mu}{\sqrt{Var[\bar{d}]/T}} \xrightarrow{d} N(0,1), \qquad \bar{d} = \frac{1}{m}\sum_{i=T+1}^{T+m} d_i$$

- Then, under H_0, the DM test is a simple *z-test*:

$$DM = \frac{\bar{d}}{\sqrt{\hat{V}ar[\bar{d}]/T}} \xrightarrow{d} N(0,1)$$

For private use only - Not to be shared/posted online without authorization.

Evaluation of forecasts: DM Test

where $\hat{V}ar[\bar{d}]$ is a consistent estimator of the variance, usually based on sample autocovariances of d_t:

$$\hat{V}ar[\bar{d}] = \gamma(0) + 2\sum_{j=k}^{\ell} \gamma(j)$$

• There are some suggestion to calculate small sample modification of the DM test. For example, :

$$\text{DM}^* = \text{DM}/\{[T + 1 - 2\,\ell + \ell\,(\ell - 1)/T]/T\}^{1/2} \sim t_{T-1}.$$

where ℓ-step ahead forecast. If time-varying volatility (ARCH) is suspected, replace ℓ with $[0.5\sqrt{(T)}] + \ell$.

Note: If $\{e_t^{M1}\}$ & $\{e_t^{M2}\}$ are perfectly correlated, the numerator and denominator of the DM test are both converging to 0 as $T \to \infty$.

⇒ Avoid DM test when this situation is suspected (say, two nested models.) Though, in small samples, it is OK.

Evaluation of forecasts: DM Test

Example: Code in R

```
 dm.test <- function (e1, e2, h = 1, power = 2) {
d <- c(abs(e1))^power - c(abs(e2))^power
 d.cov <- acf(d, na.action = na.omit, lag.max = h - 1, type = "covariance", plot = FALSE)$acf[, , 1]
 d.var <- sum(c(d.cov[1], 2 * d.cov[-1]))/length(d)
 dv <- d.var                                   #max(1e-8,d.var)
 if(dv > 0)
   STATISTIC <- mean(d, na.rm = TRUE) / sqrt(dv)
 else if(h==1)
   stop("Variance of DM statistic is zero")
 else
 {
   warning("Variance is negative, using horizon h=1")
   return(dm.test(e1,e2,alternative,h=1,power))
 }
   n <- length(d)
 k <- ((n + 1 - 2*h + (h/n) * (h-1))/n)^(1/2)
 STATISTIC <- STATISTIC * k
 names(STATISTIC) <- "DM"
}
```

For private use only - Not to be shared/posted online without authorization.

Evaluation of forecasts: DM Test

Example: We compare the SES and HW forecasts for the log of U.S. monthly vehicle sales. We use the *dm.test* function, part of the forecast package.

```
library(forecast)
> dm.test(f_error_c_ses, f_error_c_hw, power=2)

        Diebold-Mariano Test

data:  f_error_c_sesf_error_c_hw
DM = 1.6756, Forecast horizon = 1, Loss function power = 2, p-value = 0.1068
alternative hypothesis: two.sided

> dm.test(f_error_c_ses,f_error_c_hw, power=1)

        Diebold-Mariano Test

data:  f_error_c_sesf_error_c_hw
DM = 1.94, Forecast horizon = 1, Loss function power = 1, p-value = 0.064
alternative hypothesis: two.sided
```

Note: Cannot reject H_0: $MSE_{SES} = MSE_{HW}$ at 5% level

Evaluation of forecasts: DM Test – Remarks

• The DM tests is routinely used. Its "model-free" approach has appeal. There are model-dependent tests, see West (1996), Clark and McCracken (2001), and, more recent, Clark and McCracken (2011), with more complicated asymptotic distributions.

• The loss function does not need to be symmetric (like MSE).

• The DM test is based on the notion of unconditional –i.e., on average over the whole sample- expected loss.

• Following Morgan, Granger and Newbold (1977), the DM statistic can be calculated by regression of d_t, on an intercept, using NW SE. But, we can also condition on variables that may explain d_t. We move from an unconditional to a conditional expected loss perspective.

For private use only - Not to be shared/posted online without authorization.

Evaluation of forecasts – Conditional Test

• Giacomini and White (2006) present a general framework for out-of-sample predictive ability testing, characterized by the formulation of tests (such as tests for equality of forecasts) based on conditional expected loss. Now,

$$E[\hat{d}_t | I_T] = 0 \Rightarrow E[h_{t-1}\hat{d}_t] = 0.$$

where h_{t-1} is a I_T, measurable function of dimension q.

Note: G&W (2006) also differs from the standard approach to testing for predictive ability in that it compares forecasting methods (estimation + model) rather than forecasting models.

• The test becomes a Wald test, with an asymptotic χ_q^2 distribution.

Combination of Forecasts: Introduction

• Idea – from Bates & Granger (*Operations Research Quarterly,* 1969):

- We have different forecasts from R models:

$$\hat{Y}_T^{M1}(\ell), \hat{Y}_T^{M2}(\ell), \quad \ldots, \hat{Y}_T^{MR}(\ell)$$

• Instead of using the single "best model," why not combine them?

$$\hat{Y}_T^{Comb}(\ell) = \omega_{M1}\hat{Y}_T^{M1}(\ell) + \omega_{M2}\hat{Y}_T^{M2}(\ell) + \ldots + \omega_{MR}\hat{Y}_T^{MR}(\ell)$$

• $\hat{Y}_T^{Comb}(\ell)$ is usually referred as "**ensemble forecast**" or "**combination forecast**."

• Very common practice in economics, finance and politics, reported by the press as "consensus forecast." Usually, as a simple average.

• There is a strong evidence in favor of combination forecasts.

66

For private use only - Not to be shared/posted online without authorization.

Combination of Forecasts: Introduction

• Forecasts combinations have appeared in diverse areas such as retail (Ma and Fildes (2021)), energy (Xie and Hong (2016)), economics (Aastveit et al. (2019)), epidemiology (Ray et al. (2022)), etc.

• Many explanations for this strong performance:

- Incomplete information. Combining forecasts expands the information set of the individual forecasts, which are each based on partial information sets (say, private information) or models.
- Structural breaks and other instabilities. Combining forecasts from models with different degrees of misspecification and adaptability can mitigate the problem, -see Timmermann (2006) and Rossi (2021).
- Shrinkage. The unknown future value, a "meta parameter," can be improved as an average of individual estimates –see Hendry and Clements (2004).

67

Combination of Forecasts: Introduction

• The gains from forecast combinations rely on not only the quality of the individual forecasts to be combined, but the estimation of the combination weights assigned to each forecast -Cang and Yu (2014).

• Thus, forecast combinations can be linear or nonlinear, static or time-varying, series-specific or cross-learning, and ignore or cover correlations among individual forecasts.

• Mean or Median? (Or trimmed/winsorized means?) McNees (1992) found no big difference; Stock and Watson (2004) favor the mean; Jose and Winkler (2008) suggest the trimmed/winsorized means.

Note: In the 2020 M4 forecasting competition (100,000 times series & 61 methods), the simple average finished 3rd for annual time series and the median 5th for point forecasts.

68

For private use only - Not to be shared/posted online without authorization.

Combination of Forecasts: Optimal Weights

• We expect $\hat{Y}_T^{Comb}(\ell)$ to have a lower forecast variance. Why? Diversification argument. The variance of the ensemble forecast is:

$$Var[\hat{Y}_T^{Comb}(\ell)] = \sum_{j=1}^{R} (\omega_{Mj})^2 Var[\hat{Y}_T^{Mj}(\ell)] +$$
$$+ 2 \sum_{j=1}^{R} \sum_{i=j+1}^{R} \omega_{Mj}\omega_{Mi} \text{Covar}[\hat{Y}_T^{Mj}(\ell)\hat{Y}_T^{Mi}(\ell)]$$

Note: Ideally, we would like to have negatively correlated forecasts.

• Assuming unbiased forecasts and uncorrelated errors,

$$Var[\hat{Y}_T^{Comb}(\ell)] = \sum_{j=1}^{R} (\omega_{Mj})^2 \sigma_j^2$$

Example: Simple average: $\omega_j = 1/R$. Then,

$$Var[\hat{Y}_T^{Comb}(\ell)] = 1/R^2 \sum_{j=1}^{R} \sigma_j^2 .$$

Combination of Forecasts: Optimal Weights

Example: We combine the SES and HW forecast of log US vehicles sales:

```
f_comb <- (ses_f_c + car_f_hw)/2
f_error_comb <- f_comb - y[T1:T]
> var(f_comb)
[1] 0.0178981
> var(car_f_hw)
[1] 0.02042458
> var(ses_f_c)
[1] 0.01823237
```

For private use only - Not to be shared/posted online without authorization.

Combination of Forecasts: Optimal Weights

• We can derived optimal weights –i,e., ω_j's that minimize the variance of the forecast (MSE loss function). Under the uncorrelated assumption:

$$\omega_{Mj}^{*} = \sigma_j^{-2} / \sum_{j=1}^{R} \sigma_j^{-2}$$

The ω_j^{*}'s are inversely proportional to their variances.

• In general, forecasts are biased and correlated. The correlations will appear in the above formula for the optimal weights. For the two forecasts case:

$$\omega_{Mj}^{*} = (\sigma_1^2 - \sigma_{12})/(\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}) = (\sigma_1^2 - \rho\sigma_1\sigma_2)/(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)$$

• Ideally, we would like to have negatively correlated forecasts.

<u>Note</u>: Different loss functions produce different "optimal weights."

Combination of Forecasts: Regression Weights

• Granger and Ramanathan (1984) used a regression method to combine R forecasts.

- Regress the actual value on the forecasts. The estimated coefficients are the weights.

$$y_{T+\ell} = \beta_1 \hat{Y}_T^{M1}(\ell) + \beta_2 \hat{Y}_T^{M2}(\ell) + \ldots + \beta_R \hat{Y}_T^{MR}(\ell) + \varepsilon_{T+\ell}$$

• Should use a constrained regression

– Omit the constant

– Enforce non-negative coefficients.

– Constrain coefficients to sum to one.

<u>Note</u>: When $R > T$ other methods have to be used.

72

Combination of Forecasts: Regression Weights

Example: We regress the SES and HW forecasts against the observed car sales to obtain optimal weights. We omit the constant

```
> lm(y[T1:T] ~ ses_f_c + car_f_hw - 1)

Call:
lm(formula = y[T1:T] ~ ses_f_c + car_f_hw - 1)

Coefficients:
 ses_f_c  car_f_hw
 -0.5426    1.5472
```

Note: Coefficients (weights) add up to 1. But, we see negative weights... In general, we use a constrained regression, forcing parameters to be between 0 and 1 (& non-negative). But, h=25 delivers not a lot of observations to do non-linear estimation.

Combination of Forecasts: Regression Weights

• Chan et al. (1999) show that OLS combinations have poor performance when R is very large. Principal components regression (PCR) can be used, resulting in a two-step procedure:

(1) Extracts the principal components

(2) Use PC to forecast using OLS regression.

• Rapach and Strauss (2008) and Poncela et al. (2011) find better PCR performance than OLS.

• Q: Should we forecast with variables (competing point forecasts), factors (extracted from the R competing forecasts), or both -see, Castle et al. (2013).

74

For private use only - Not to be shared/posted online without authorization.

Combination of Forecasts: Regression Weights

• Remarks:

- To get weights, do not include a constant. Here, we are assuming unbiased forecasts. If the forecasts are biased, we include a constant.
- To account for potential correlation of errors, we can allow for ARMA residuals or include $y_{T+\ell+1}$ in the regression.
- Time varying weights are also possible –see Deutsch et al. (1994).

• Many methods to get weights: Bayesian, IC, Historical, ML, etc.

• Should weights matter? Two views:

- Simple averages outperform more complicated combination techniques. Stock and Watson (2004), Chan and Pauwels (2018)
- Sampling variability may affect weight estimates to the extent that the combination has a larger MSE.

75

Combination of Forecasts: Bayesian Weights

• In our discussion of model selection, we mentioned that the BIC is consistent. That means, the probability that a model is true, given the data is proportional to BIC:

$$\text{P}(M_j|\text{data}) \propto \exp(-\frac{BIC_j}{2}).$$

• Based on this, we use the BIC of different models to derive weights. This is a simplified form of **Bayesian model averaging** (**BMA**).

• Easy calculation of weights. Let $BIC*$ be the smallest BIC among the R models considered. Define $\Delta BIC_{M_j} = BIC_{M_j} - BIC*$.

Then,
$$\omega_{M_j}{}^* = \exp(-\frac{\Delta BIC_{M_j}}{2}).$$

$$\omega_{M_j} = \frac{\omega_{M_j}{}^*}{\sum_{j=1}^{R} \omega_{M_j}{}^*}$$

76

For private use only - Not to be shared/posted online without authorization.

Combination of Forecasts: Bayesian Weights

• Steps:

(1) Compute BIC for the R different models.

(2) Find best-fitting $BIC*$.

(3) Compute ΔBIC & $\exp(-\Delta BIC/2)$.

(4) Add up all values and re-normalize.

• BMA puts the most weight on the model with the smallest BIC.

• Some authors have suggested replacing BIC with AIC in the weight formula –i.e., $\omega_j \propto \exp(-\frac{AIC_j}{2})$.

- There is no clear theory for this formula. It is simple and works well in practice.

- This method is called **weighted AIC** ($WAIC$).

77

Combination of Forecasts: Bayesian Weights

• Q: Does it make a difference the criteria used? Two situations:

(1) The selection criterion (AIC, BIC) are close for competing models. Then, it is difficult to select one over the other.

- $WAIC$ and BMA will produce similar weights.

(2) The selection criterion are different.

- $WAIC$ and BMA will produce different weights.
- They will give zero weight if the difference is large, say, above 10.

Q: Which one to use?

- Not clear. $WAIC$ works well in practice.

General finding: Simple averaging works well, but it is not optimal. A combination beats the lowest criteria used.

78

For private use only - Not to be shared/posted online without authorization.

Combination of Forecasts: Final Comments

• A simple average, with equal weights, tends to do well ("*forecast combination puzzle*".) However, there is a large "optimal weights" literature.

• Traditionally, optimal combination weights have generally been chosen to minimize a symmetric, squared-error loss function.

• But, asymmetric loss functions can also be used. Elliot and Timmermann (2004) allow for general loss functions (and distributions). They find that the optimal weights depend on higher order moments, such a skewness.

• Ideally, an increase in diversity among forecasting models has the potential to improve the accuracy of their combination. We prefer forecasts with low correlation (higher diversity).

79

Combination of Forecasts: Final Comments

• Non-linear combinations are possible, for example, using ML -see Krasnopolsky and Lin (2012) and Babikir and Mwambi (2016), used neural networks (ANNs).

• There is a literature developing a set of rules and features to be used to combine forecasts –Collopy and Armstrong (1992), Petropoulos et al. (2014). There is an R package (FFORMA) implementing some of the rules and features (it finished 2nd in the M4 competition).

• A big literature on combining **probability forecasts**. For example, forecast quantiles and combine them through averaging –see Busetti (2017). Testing of quantile forecasts can be based on the general approach of G&W (2006). Giacomini and Komunjer (2005) present an application.

80

For private use only - Not to be shared/posted online without authorization.