

© R. Susmel, 2023 (for private use, not to be posted/shared online).







### **Removing Seasonal Patterns**

• In the presence of seasonal patterns, we proceed to do seasonal adjustments to remove these predictable influences.

• Seasonalities can blur both the true underlying movement in the series, as well as certain non-seasonal characteristics which may be of interest to analysts.

• Similar to the trend, the type of adjustment depends on how we view the seasonal pattern: **Deterministic** or **Stochastic**.

• In this class, we focus on the deterministic seasonal pattern, which can be **additive** (constant pattern) or **multiplicative** (increasing pattern over time).

### **Removing Seasonal Patterns**

• We follow a similar 2-step process to detrending:

- 1) Regress  $y_t$  against the seasonal dummies. Keep residuals
- 2) With the residuals, follow Box-Jenkins to select an ARIMA model.

• For Step 1. Suppose  $y_t$  has monthly frequency, we suspect that  $y_t$  increases every December.

– For the **additive model**, we regress  $y_t$  against a constant and a December dummy,  $D_t$ :

 $y_t = \mu + \boldsymbol{D}_t \boldsymbol{\mu}_s + \varepsilon_t$ 

– For the **multiplicative model**, we regress  $y_t$  against a constant and a December dummy,  $D_t$ , interacting with a trend:

$$y_t = \mu + \boldsymbol{D}_t \boldsymbol{\mu}_s * t + \varepsilon_t$$

• For Step 2. Use the residuals of these regressions,  $e_t$ , –i.e.,  $e_t = filtered y_t$ , free of "monthly seasonal effects"– for ARMA modeling.







x\_la\_fit\_sea <- lm(x\_la ~ seas) > summary(x\_la\_fit\_sea) # Regress x\_la against constant + seasonal dummies







## Non-Stationarity in Variance: Logs

• When we work with a nominal series (not changes, say, USD total retail sales or total units sold), it is common to first apply a *variance stabilizing transformation* to the data, usually using logs.

• Many times, this stabilizing transformation is done because the variance is non-stationary. In practice, a variance stabilizing transformation is done to reduce the variance of the series.

• In the case of non-stationary variance, we want to find a function G(.) so that the transformed series  $G(y_t)$  has a constant variance.

• Very popular transformation:  $G(Y_t) = \log(Y_t)$ 





## Forecasting (Again)

• Forecasting is the primary objective of ARIMA modeling. We focus on **out-of-sample** forecasting: Produce an estimate of a future value, outside the sample,  $\hat{Y}_{T+\ell}$ .

• Forecast: Conditional expectation of  $Y_{T+\ell}$ , given  $I_T$ :  $\hat{Y}_{T+\ell} = E[Y_{T+\ell} | I_T = \{Y_T, Y_{T-1}, ..., Y_1, \varepsilon_1, \varepsilon_2, ..., \varepsilon_T\}]$ 

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:  $e_{T+\ell} = e_T(\ell) = Y_{T+\ell} - \hat{Y}_{T+\ell}$ - Mean squared error (MSE):  $MSE(e_{T+\ell}) = E[Y_{T+\ell} - \hat{Y}_{T+\ell}]^2$ <u>Remark</u>: Under MSE, the optimal forecast is the (conditional) mean:  $\hat{Y}_{T+\ell} = E[Y_{T+\ell} | I_T]$ 



### Forecasting From ARIMA 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  $\mu$ ,  $\phi_1$  and  $\phi_2$ :  $\hat{\mu}$ ,  $\hat{\phi}_1 & \hat{\phi}_2$ . (3) We find residuals are WN. (4) Now, we forecast. The one-step ahead forecast at time T:  $\hat{Y}_{T+1} = E[Y_{T+1} | I_T = \{Y_T, Y_{T-1}, ..., Y_1\}] = \hat{\mu} + \hat{\phi}_1 Y_T + \hat{\phi}_2 Y_{T-1}$ At time T + 1, we compute the one-step ahead forecast error,  $e_T(1)$ :  $e_T(1) = Y_{T+1} - \hat{Y}_{T+1}$ <u>Note</u>: After Q periods, we compute Q one-step ahead forecast errors and MSE.

### Forecasting From MA(q) Models

• The stationary MA(q) model for  $Y_t$  is  $Y_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q}$ 

We produce at time *T l*-step ahead forecasts using:

$$Y_{T+1} = \mu + \varepsilon_{T+1} + \theta_1 \varepsilon_T + \dots + \theta_q \varepsilon_{T-q+1}$$

$$Y_{T+2} = \mu + \varepsilon_{T+2} + \theta_1 \varepsilon_{T+1} + \dots + \theta_q \varepsilon_{T-q+2}$$

$$\vdots$$

$$Y_{T+\ell} = \mu + \varepsilon_{T+l} + \theta_1 \varepsilon_{T+l-1} + \dots + \theta_q \varepsilon_{T+l-q} \qquad (l > 2)$$

Now, we take conditional expectations:

$$\hat{Y}_{T+\ell} = E[Y_{T+\ell} | I_T] = \mu + E[\varepsilon_{T+\ell} | I_T] + \theta_1 E[\varepsilon_{T+\ell-1} | I_T] + \dots + \theta_q E[\varepsilon_{T+\ell-q} | I_T]$$

Note: Forecasts are a linear combination of errors.

### Forecasting From MA(q) Models

• Some of the errors are know at  $T: \varepsilon_1 = \widehat{\varepsilon}_1, \varepsilon_2 = \widehat{\varepsilon}_2, ..., \varepsilon_T = \widehat{\varepsilon}_T$ , the rest are unknown. Thus,  $E[\varepsilon_{T+j}] = 0 \qquad \text{for } j > 1.$ Example: For an MA(2) we have:  $\widehat{Y}_{T+1} = \mu + E[\varepsilon_{T+1} | I_T] + \theta_1 E[\varepsilon_T | I_T] + \theta_2 E[\varepsilon_{T-1} | I_T]$   $\widehat{Y}_{T+2} = \mu + E[\varepsilon_{T+2} | I_T] + \theta_1 E[\varepsilon_{T+1} | I_T] + \theta_2 E[\varepsilon_T | I_T]$   $\widehat{Y}_{T+3} = \mu + E[\varepsilon_{T+3} | I_T] + \theta_1 E[\varepsilon_{T+2} | I_T] + \theta_2 E[\varepsilon_{T+1} | I_T]$ At time T = t, we know  $\varepsilon_t \& \varepsilon_{t-1}$ . Set  $E[\varepsilon_{t+j} | I_j] = 0$  for j > 1. Then,  $\widehat{Y}_{t+1} = \mu + \theta_1 E[\varepsilon_t | I_j] + \theta_2 E[\varepsilon_{t-1} | I_t] = \mu + \theta_1 \widehat{\varepsilon}_t + \theta_2 \widehat{\varepsilon}_{t-1}$   $\widehat{Y}_{t+2} = \mu + \theta_2 E[\varepsilon_t | I_t] = \mu + \theta_2 \widehat{\varepsilon}_t$   $\widehat{Y}_{t+3} = \mu$  $\widehat{Y}_{t+4} = \mu$  for  $\ell > 2$ .  $\Rightarrow$  MA(2) memory of 2 periods





### Forecasting From AR(p) Models

**Example:** AR(2) model for  $Y_{t+\ell}$  is  $Y_{t+\ell} = \mu + \phi_1 Y_{t+\ell-1} + \phi_2 Y_{t+\ell-2} + \varepsilon_{t+\ell}$ 

Then, taking conditional expectations at T = t, we get the forecasts:

$$\begin{split} \hat{Y}_{t+1} &= \mu + \phi_1 Y_t + \phi_2 Y_{t-1} \\ \hat{Y}_{t+2} &= \mu + \phi_1 \hat{Y}_{t+1} + \phi_2 Y_t \\ \hat{Y}_{t+3} &= \mu + \phi_1 \hat{Y}_{t+2} + \phi_2 \hat{Y}_{t+1} \\ \vdots \\ \hat{Y}_{t+\ell} &= \mu + \phi_1 \hat{Y}_{t+\ell-1} + \phi_2 \hat{Y}_{T+\ell-1} \end{split}$$

• AR-based forecasts are autocorrelated, they have long memory!

• The forecast combine (& recombine) the observed data  $Y_T \& Y_{T-1}$ :  $\hat{Y}_{T+2} = \mu + \phi_1(\mu + \phi_1 Y_T + \phi_2 Y_{T-1}) + \phi_2 Y_T$  $= \mu (1 + \phi_1) + (\phi_1^2 + \phi_2) Y_T + \phi_1 \phi_2 Y_{T-1}$ 

### Forecasting From AR(p) Models

• The forecasts combine (& recombine) the observed data  $Y_T \& Y_{T-1}$ . For example, for  $\hat{Y}_{T+2}$ :  $\hat{Y}_{T+2} = \mu + \phi_1(\mu + \phi_1Y_T + \phi_2Y_{T-1}) + \phi_2Y_T$   $= \mu (1 + \phi_1) + (\phi_1^2 + \phi_2)Y_T + \phi_1\phi_2Y_{T-1}$ Similarly for  $\hat{Y}_{T+3}$ :  $\hat{Y}_{T+3} = \mu + \phi_1\hat{Y}_{T+2} + \phi_2\hat{Y}_{T+1}$   $= \mu + \phi_1(\mu (1 + \phi_1) + (\phi_1^2 + \phi_2)Y_T + \phi_1\phi_2Y_{T-1})$   $+ \phi_2(\mu + \phi_1Y_T + \phi_2Y_{T-1})$  $= \mu (1 + \phi_1 + \phi_2 + \phi_1^2) + (\phi_1^3 + 2\phi_2\phi_1)Y_T + (\phi_1^2\phi_2 + \phi_2^2)Y_{T-1}$ 



# **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. $\hat{Y}_{T+2} = \mu + \phi_1 E[Y_{T+1}|I_T] + E[\varepsilon_{T+2}|I_T] + \theta_1 E[\varepsilon_{T+1}|I_T] + \theta_2 E[\varepsilon_T|I_T]$ $= \mu + \phi_1 \hat{Y}_{t+1} + \theta_2 \hat{\varepsilon}_T$ Actual: $Y_{T+2} = \mu + \phi_1 Y_{T+1} + \varepsilon_{T+2} + \theta_1 \varepsilon_{T+1} + \theta_2 \hat{\varepsilon}_T$ $e_T(2) = Y_{T+2} - \hat{Y}_{T+2} = \phi_1(\hat{Y}_{t+1} - Y_{T+1}) + \varepsilon_{T+2} + \theta_1 \varepsilon_{T+1}$

# 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: $Y_{T+\ell} = \mu + \varepsilon_{T+\ell} + \Psi_1\varepsilon_{T+\ell-1} + \Psi_2\varepsilon_{T+\ell-2} + \dots + \Psi_\ell \varepsilon_T + \dots$ • The Wold representation depends on an infinite number of parameters, but, in practice, they decay rapidly. • The forecast error is: $e_T(\ell) = \sum_{i=0}^{\ell-1} \Psi_i \varepsilon_{T+\ell-i}$ ( $\Psi_0 = 1$ ) Note: If $E[e_T(\ell)] = 0$ , we say the forecast is unbiased.

### Forecasting From ARMA Models

• The forecast erro <i>e<sub>T</sub></i>	or is: $(\ell) = \sum_{i=0}^{\ell-1} \Psi_i \varepsilon_{T+\ell-i}$	$(\Psi_0 = 1)$
• The variance of $Var(e_T(\ell)) =$	the forecast error: = $Var(\sum_{i=0}^{\ell-1} \Psi_i \varepsilon_{T+\ell-i}) = \sigma^2$	$\sum_{i=0}^{\ell-1} \Psi_i^2 \qquad (\Psi_0 = 1)$
<b>Example:</b> One-st $Y_{T+1} = \mu$	ep ahead forecast ( $\ell = 1$ ). + $\varepsilon_{T+1} + \Psi_1 \varepsilon_T + \Psi_2 \varepsilon_{T-1} +$	$\Psi_3 \varepsilon_{T-2} + \cdots$
Forecast:	$\hat{Y}_{T+1} = \mu + \Psi_1 \varepsilon_T + \Psi_2 \varepsilon_T$	$\varepsilon_{T-1} + \cdots$
Forecast error:	$e_T(1) = Y_{T+1} - \hat{Y}_{T+1} =$	$\varepsilon_{T+1}$
Variance:	$Var(e_T(1)) = \sigma^2$	
For the two-step a	thead forecast ( $\ell = 2$ ). $e_T(2) = Y_{T+2} - \hat{Y}_{T+2} =$	$\varepsilon_{E_{T+2}} + \Psi_{1}\varepsilon_{T+4}$
	$Var(e_T(2)) = \sigma^2 * (1 - 1)$	$+\Psi_1^2$ ) $+\Psi_1^2$ )

## Forecasting From ARMA Models

• In the Wold representation, in practice, the parameters,  $\Psi_i$ 's, decay rapidly. Then, as we forecast into the future, the forecasts tend to the unconditional forecasts,  $\mu$  and  $\sigma^2$ :

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

Not very interesting.

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

### Forecasting From ARMA Models: C.I.

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

$$\hat{Y}_{T}(\ell) \pm z_{\alpha/2} \sqrt{Var(e_{T}(\ell))}$$
$$\hat{Y}_{T}(\ell) \pm z_{\alpha/2} \sigma \sqrt{\sum_{i=0}^{\ell-1} \Psi_{i}^{2}}$$

or,

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

## 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**).

### Simple Exponential Smoothing: SES

• We "**smooth**" the series  $Y_t$  to produce a quick forecast,  $S_{t+1}$ , also called *level's forecast*. Smooth? The graph of  $S_t$  is less jagged than the graph of the original series,  $Y_t$ .

• We use the observed time series at time  $t : Y_1, Y_2, ..., Y_t$ .

• The equation for the **level**:  $S_t = \alpha Y_{t-1} + (1 - \alpha)S_{t-1}$ where

-  $\alpha$ : The smoothing parameter,  $0 \le \alpha \le 1$ .

-  $Y_t$ : Value of the observation at time t.

-  $S_t$ : Value of the smoothed observation at time t –i.e., the forecast.

• The equation can also be written as an **updating equation**:

 $S_t = S_{t-1} + \alpha \left( Y_{t-1} - S_{t-1} \right) = S_{t-1} + \alpha * \text{(past forecast error)}$ 

### SES: Forecast and Updating

• From the updating equation *S<sub>t</sub>*:

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

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

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

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  $\ell$ -step ahead forecast is:

 $S_{t+\ell} = S_{t+1} \implies$  A naive forecast!

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

### 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  $\alpha = 0.25$ . The firm observes  $Y_t = 5$  and, last period's forecast,  $S_t = 3$ . Then, the forecast for time t + 1 is:  $S_{t+1} = 3 + 0.25 * (5 - 3) = 3.50$ The forecast for time t + 1 (& any period after time t + 1) is:  $S_{t+\ell} = S_{t+1} = 3.50$  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:  $MSE = \frac{1}{3} * [(4.77 - 3.50)^2 + (3.15 - 3.50)^2 + (1.85 - 3.50)^2] = 1.486$ .

## SES: Forecast and Updating

Example (continuation):Note: If  $\alpha = 0.75$ , then $S_{t+1} = 3 + 0.75 * (5 - 3) = 4.50$ A bigger  $\alpha$  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: $S_{t+\ell} = S_{t+1} = 4.50$  for  $\ell > 1$ .

### SES: Exponential?

• Q: Why Exponential?

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

$$\begin{split} S_{t+1} &= \alpha Y_t + (1-\alpha)S_t = \alpha Y_t + (1-\alpha)[\alpha Y_{t-1} + (1-\alpha)S_{t-1}] \\ &= \alpha Y_t + \alpha (1-\alpha)Y_{t-1} + (1-\alpha)^2 S_{t-1} \end{split}$$

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

where  $c_i$ 's are the weights, with

 $c_i = \alpha (1-\alpha)^i; i = 0, 1, \dots; 0 \le \alpha \le 1.$ 

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

36

SES: Exponential Weights					
• $c_i = \alpha (1 - \alpha)^i$ ; $i = 0, 1,; 0 \le \alpha \le 1$ .					
$c_i = \alpha (1-\alpha)^i$	$\alpha = 0.25$	<i>α</i> = 0.75			
<i>c</i> <sub>0</sub>	0.25	0.75			
<i>c</i> <sub>1</sub>	0.25 * 0.75 = 0.1875	0.75 * 0.25 = 0.1875			
<i>c</i> <sub>2</sub>	$.25 * 0.75^2 = 0.140625$	$0.75 * 0.25^2 = 0.046875$			
<i>c</i> <sub>3</sub>	$.25 * 0.75^3 = 0.1054688$	$0.75 * 0.25^3 = 0.01171875$			
<i>C</i> <sub>4</sub>	$.25 * 0.75^4 = 0.07910156$	$0.75 * 0.25^4 = 0.002929688$			
:					
<i>c</i> <sub>12</sub>	$.25 * 0.75^{12} = 0.007919088$	$0.75 * 0.25^{12} = 4.470348e-08$			
	1	1			

Decaying weights. Faster decay with greater  $\alpha$ , associated with faster learning: we give more weight to more recent observations.

• We do not know  $\alpha$ ; we need to estimate it.

37

## SES: Selecting $\alpha$

• Choose  $\alpha$  between 0 and 1.

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

- If  $\alpha$  is close to 0, distant values are given weights comparable to recent values. Set  $\alpha \approx 0$  when there are big random variations in  $Y_t$ . -  $\alpha$  is often selected as to minimize the MSE.

• In empirical work,  $0.05 \le \alpha \le 0.3$  are used ( $\alpha \approx 1$  is used rarely).

Numerical Minimization Process:

- Take different  $\alpha$  values ranging between 0 and 1.
- Calculate 1-step-ahead forecast errors for each  $\alpha$ .
- Calculate MSE for each case.

Choose  $\alpha$  which has the min MSE:  $e_t = Y_t - S_t \Rightarrow \min \sum_{t=1}^n e_t^2 \stackrel{_{38}}{\Rightarrow} \alpha$ 

$S_{t+1} = \alpha Y_t + (1-\alpha)S_t$				
Time	$Y_t$	$S_{t+1}(\alpha = 0.10)$	$(Y_t - S_t)^2$	
1	5	-	-	
2	7	(0.1) <b>5</b> + (0.9) <b>5</b> = <b>5</b>	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	
5	4 <i>MSE</i> =	(0.1)3 + (0.9)5.28 = 5.052 TOTAL = $\frac{SSE}{1} = 2.74$	1.107 <b>10.94</b> 5	

## **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  $\alpha$ .)

40







Example 1 (continuation): Not	w, we do one-step ahead forecasts
T_last <- nrow(mod1\$fitted)	# number of in-sample forecasts
h <- 25	# forecast horizon
$ses_f \le matrix(0,h,1)$	# Vector to collect forecasts
alpha <- 0.29	
y <- lr_d	
$T \leq - length(lr_d)$	
$sm \leq -matrix(0,T,1)$	
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 $(a \le T)$ {	
sm[a] = alpha * y[a-1] + (1-alpha)	u) * 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")









## SES: Remarks

• Some computer programs automatically select the optimal  $\alpha$ , 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<sub>t</sub>, can be easily incorporated.

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

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

<u>Note</u>: Seasonal factor is multiplied in the h-step ahead forecast.



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

Holt-Winters (HW) ES: Additive					
• Then, the n	ullet Then, the model for the $h$ -step ahead forecast				
$\hat{Y}_t(h) = S_t + h T_t + I_{t+h-s}$					
has three equ	ations:				
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 or $\alpha = 1$ $\beta = t$	ly three smoothing parameters: evel coefficient rend coefficient				
$\gamma$ = seasonality coefficient					

## Holt-Winters (HW) ES: Multiplicative

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

$$Y_t(h) = (S_t + h T_t) * I_{t+h-s}$$

• 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:

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

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

$$I_{t} = \gamma \frac{Y_{t}}{S_{t}} + (1 - \gamma) I_{t-s}$$

54



### 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:  $\alpha = 0.25; \beta = 0.1; \& \gamma = 0.4$ . 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:  $S_t = 0.25 \frac{5}{1.1} + (1 - 0.25) * (3 + 1.3) = 4.2864$  Holt-Winters (HW) ES: Multiplicative Example (continuation):  $(\alpha = 0.25; \beta = 0.1; \& \gamma = 0.4.)$   $S_t = 0.25 * \frac{5}{1.1} + (1 - 0.25) * (3 + 1.2) = 4.2864$   $T_t = 0.1 * (4.2864 - 3) + (1 - 0.1) * 1.2 = 1.2086$   $I_t = 0.4 * \frac{5}{4.2864} + (1 - 0.4) * 1.1 = 1.1266$ 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:

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

$$T_{0} = \frac{1}{s} \left( \frac{Y_{s+1} - Y_{1}}{s} + \frac{Y_{s+2} - Y_{2}}{s} + \dots + \frac{Y_{s+s} - Y_{s}}{s} \right)$$
  
or  $T_{0} = \left[ \left\{ \sum_{t=1}^{s} Y_{t}/s \right\} - \left\{ \sum_{t=s+1}^{2s} Y_{t}/s \right\} \right] / s$ 

58

### HW ES: Initial Values

Algorithm to compute initial values for seasonal component I<sub>s</sub>. Assume we have *T* observation and quarterly seasonality (*s*=4):
(1) Compute the averages of each of *T* years.
A<sub>t</sub> = ∑<sup>4</sup><sub>t=1</sub> Y<sub>t,i</sub>/4, t = 1, 2, ..., 6 (yearly averages)
(2) Divide the observations by the appropriate yearly mean: Y<sub>t,i</sub>/A<sub>t</sub>.
(3) I<sub>s</sub> is formed by computing the average Y<sub>t,i</sub>/A<sub>t</sub> per year:
I<sub>s</sub> = ∑<sup>T</sup><sub>i=1</sub> Y<sub>t,s</sub>/A<sub>t</sub> s = 1, 2, 3, 4

### HW ES: Damped Model

• We can damp the trend as the forecast horizon increases, using a parameter  $\phi$ . 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:  $\hat{Y}_t(h) = \{S_t + (1 + \phi + \phi^2 + \dots + \phi^{2h-1})T_t\} * I_{t+h-s}$ 

• 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.















## **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:  $Forecast_t = S_t + I_{t-s}$ .

© R. Susmel, 2023 – Not to be posted/shared online without written authorization

## Evaluation of forecasts – Accuracy measures

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

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

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) =  $\frac{100}{T (m-1)} \sum_{i=T+1}^{T+m} \left| \frac{\hat{y}_i y_i}{\hat{y}_i + y_i} \right|$

<u>Remark</u>: 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. <u>Remark</u>: 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.





### 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:  $g(e_t) = \exp(\lambda e_t^2) \lambda e_t^2$  ( $\lambda \ge 0$ ).

<u>Note</u>: 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  $\ell$  periods) needed to apply CLT. Then,

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

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

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

### Evaluation of forecasts - DM Test

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

$$\widehat{V}ar[\overline{d}] = \gamma(0) + 2\sum_{j=k}^{r} \gamma(j)$$

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

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

where  $\ell$ -step ahead forecast. If time-varying volatility (ARCH) is suspected, replace  $\ell$  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  $T \to \infty$ .  $\Rightarrow$  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)) } $n \leq - length(d)$ $k \le ((n + 1 - 2*h + (h/n) * (h-1))/n)^{(1/2)}$ STATISTIC <- STATISTIC \* k names(STATISTIC) <- "DM"



**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
```

<u>Note</u>: Cannot reject  $H_0$ : MSE<sub>SES</sub> = MSE<sub>HW</sub> at 5% level



### **Combination of Forecasts**

- Idea from Bates & Granger (Operations Research Quarterly, 1969):
- We have different forecasts from R models:

 $\hat{Y}_T^{M1}(\ell), \hat{Y}_T^{M2}(\ell), \qquad \dots, \hat{Y}_T^{MR}(\ell)$ 

• Q: Why not combine them?

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

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

• Q: Advantage? Lower forecast variance. Diversification argument.

Intuition: Individual forecasts are each based on partial information sets (say, private information) or models.

# Combination of Forecasts – Optimal Weights

• The variance of the forecasts is: R

$$Var[\hat{Y}_{T}^{Comb}(\ell)] = \sum_{j=1}^{r} (\omega_{Mj})^{2} Var[\hat{Y}_{T}^{Mj}(\ell)] + 2\sum_{j=1}^{r} \sum_{i=j+1}^{r} \omega_{Mj} \omega_{Mi} \operatorname{Covar}[\hat{Y}_{T}^{Mj}(\ell) \hat{Y}_{T}^{Mi}(\ell)]$$

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:  $\omega_i = 1/R$ . 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

## **Combination of Forecasts – Optimal Weights**

• We can derived optimal weights –i,e.,  $\omega_j$ 's that minimize the variance of the forecast. Under the uncorrelated assumption:

Under the uncorrelated assumption:

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

• The  $\omega_i^*$ '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:

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

## **Combination of Forecasts: Regression Weights**

• In general, forecasts are biased and correlated. The correlations will appear in the above formula for the optimal weights. Ideally, we would like to have negatively correlated forecasts.

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

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

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

• Should use a constrained regression

- Omit the constant

- Enforce non-negative coefficients.

- Constrain coefficients to sum to one

## Combination of Forecasts: Regression Weights Example: We regress the SES and HW forecasts against the

83

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

<u>Note</u>: 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**

• Remarks:

- To get weights, we 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  $y_{T+l\mbox{-}1}$  in the regression.

- Time varying weights are also possible.

• Should weights matter? Two views:

- Simple averages outperform more complicated combination techniques.

- Sampling variability may affect weight estimates to the extent that the combination has a larger MSE.

85

## **Forecasting: Final Comments**

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

• But, asymmetric loss functions can also be used. More recent research work find that the optimal weights depend on higher order moments, such a skewness.

86