Demand forecasting and modeling

Pramesh Kumar

IIT Delhi

April 23, 2025

Reading material

- Goetschalckx Chapter 3-4
- Ghiani et al. Chapter 2
- Chopra et al. Chapter 7
- Chapter 2 from Production and Operations Management by Steven Nahmias, 2018

Need for forecasting

- Firms need forecasting of sales of both new and existing products for planning and execution of all operations activities.
- Forecasting can be a source of competitive advantage by improving customer service and costs associated to mismatch between supply and demand.

Production philosophies

Definition (Make-to-order (MTO)).: Production is only started after the complete order for a product been received. This is to avoid forecasting of demand.

Definition (Make-to-stock (MTS)).: Production planning that procures in anticipation of forecasted future demand.

Outline

Characteristics of forecasts

- They are usually wrong.
- A good forecast is more than a single number.
- Aggregate forecasts are more accurate than disaggregate forecasts
- The longer the forecast horizon, the less accurate the forecast will be
- The farther up the supply chain a company is (or the farther from the customer), the greater the distortion of information it receives. Collaborative forecasting based on sales to the end consumer helps upstream enterprises reduce the forecast error.
- Forecast should not be used to the exclusion of known information.

Pattern classification

- Demand pattern (observations over time) can be regular or irregular.
- If the pattern is regular the future values can be predicted based on past or historical values.
- Common regular patterns are constant pattern, trend pattern, seasonal pattern, or combination of both trend and seasonal pattern.
- It is common to decompose any forecast into trend, seasonal, and random components.
- As long as random component is small compared to the underlying pattern, accurate forecasts can be obtained using mathematical techniques such as regression and time-series.
- Once the demand for final product is known, it is usually easy to know the demand for subcomponents, raw materials, and resources. The derived demand is recorded in bill of materials (BOM).

Definition (Trend). Demand consistently increases or decreases over time.

Definition (Seasonality). Demand shows peaks and valleys at consistent intervals.

Definition (Random error). Variations that cannot be explained or predicted.

Pattern classification

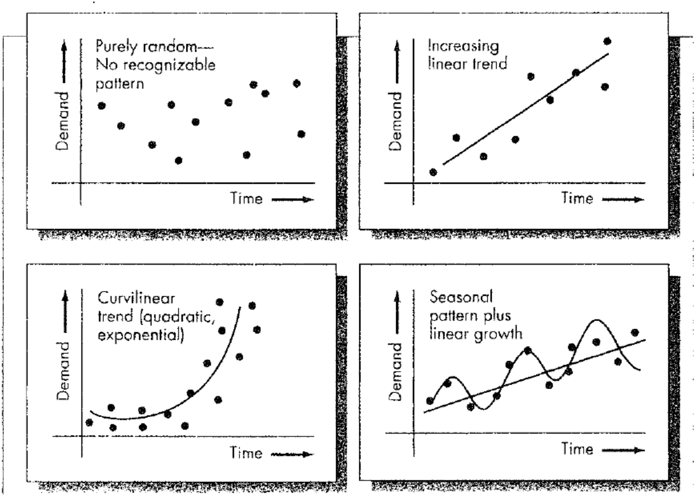


Figure: Time series patterns[1]

[1] Source: Chapter 2, Production and Operations Management by Steven Nahmias, 2018

Outline

Types of forecasting methods

1. Subjective methods
   - Sales force composites
   - Customer surveys
   - Jury of executive opinion
   - The Delphi method
2. Objective methods
   - Causal models
   - Time-series models

Notations

- D_t: observed demand for time period t
- $\{D_t\}_{t \geq 1}$: time series of observed demand
- $F_{t,t+\tau}$: forecast made in time period t for period $t + \tau$ (we are making forecast τ period into the future after observing the demand for time period t and before observing the demand for time period $t + 1$)
- $F_t = F_{t-1,t}$: forecast made in period $t - 1$ for period t (one-step-ahead forecast)

Forecast error

e_t: forecast error in period t (difference between the forecast value and actual demand for that period). Also referred to as residual.

For multiple-step-ahead forecast,

$$e_t = F_{t-\tau,t} - D_t \tag{1}$$

For one-step-ahead forecast,

$$e_t = F_t - D_t \tag{2}$$

Let $e_1, \cdots, e_n$ be the forecast errors observed over n time periods.

Error bias

A desirable property of forecasts is that they should be unbiased. Mathematically, $\mathbb{E}\{e_i\} = 0$.

FIGURE 2–3

Forecast errors over time

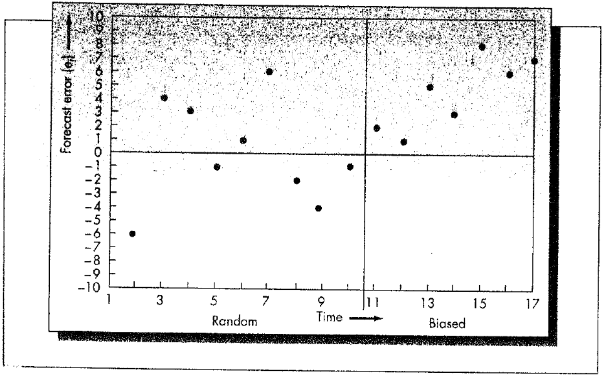


Figure: Forecast errors over time[2]

[2]Source: Chapter 2, Production and Operations Management by Steven Nahmias, 2018

Measures for evaluating forecasts

Let $e_1, \cdots, e_n$ be the forecast errors observed over n time periods.

1. Mean absolute deviation (MAD)

$$\text{MAD} = \frac{1}{n}\sum_{i=1}^{n}|e_i| \tag{3}$$

2. Mean squared error (MSE)

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}e_i^2 \tag{4}$$

3. Root mean squared error (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}e_i^2} \tag{5}$$

4. Mean absolute percentage error (MAPE)

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{e_i}{D_i}\right| \times 100 \tag{6}$$

Outline

Methods for forecasting stationary series

Definition (Stationary series). A series with each observation can be represented by a constant plus a random variation component.

$$D_t = \mu + \epsilon_t \tag{7}$$

where μ is the constant corresponding to the mean of the series and ϵ_t is random error with mean zero and variance σ^2.

We will discuss the following two methods:

1. Moving averages
2. Exponential smoothing

Moving averages

Definition (Moving average). A moving average of order N is simply the arithmetic average of the most recent N observations.

For one-step-ahead forecasting,

$$F_t = \frac{1}{N} \sum_{i=t-N}^{t-1} D_i \tag{8}$$

Remark. Multi-step-ahead and one-step-ahead forecasts are identical in case of moving averages.

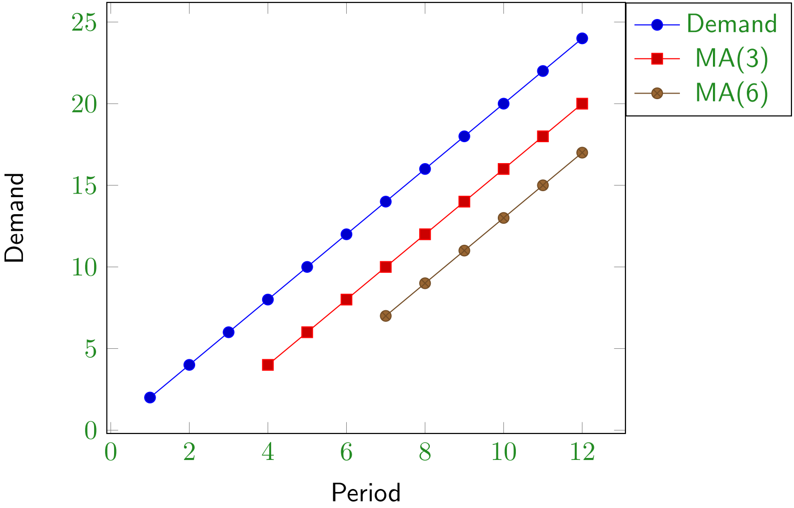
$$F_{t+\tau} = F_{t+1}, \forall \tau \geq 1 \tag{9}$$

Moving averages

Remark. Moving averages lags behind the trend. Consider the following example.

Period	Demand	MA(3)	MA(6)
1	2		
2	4		
3	6		
4	8	4	
5	10	6	
6	12	8	
7	14	10	7
8	16	12	9
9	18	14	11
10	20	16	13
11	22	18	15
12	24	20	17

Moving averages



Exponential smoothing

In this case, the current forecast is the convex combination of last forecast and the current values of demand.

$$F_t = \alpha D_{t-1} + (1-\alpha)F_{t-1} \tag{10}$$

where $0 \leq \alpha \leq 1$ is called smoothing constant. One can write,

$$F_t = F_{t-1} - \alpha(F_{t-1} - D_{t-1}) = F_{t-1} - \alpha e_{t-1} \tag{11}$$

i.e., forecast in period t is forecast in period $t-1$ minus some fraction of observed forecast error in period $t-1$.
For period $t-1$,

$$F_{t-1} = \alpha D_{t-2} + (1-\alpha)F_{t-2} \tag{12}$$

Substituting (12) in (10), we get

$$\begin{aligned} F_t =& \alpha D_{t-1} + (1-\alpha)(\alpha D_{t-2} + (1-\alpha)F_{t-2}) \\ =& \alpha D_{t-1} + \alpha(1-\alpha)D_{t-2} + (1-\alpha)^2 F_{t-2} \end{aligned}$$

Exponential smoothing

Continuing in the same fashion,

$$F_t = \sum_{i=0}^{\infty} \alpha(1-\alpha)^i D_{t-i-1} = \sum_{i=0}^{\infty} a_i D_{t-i-1}$$

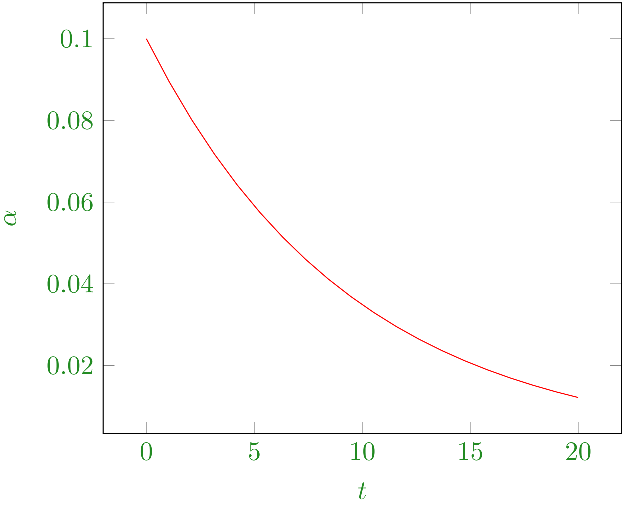
where, weights are $a_0 > a_1 > \cdots > a_i = \alpha(1-\alpha)^i$ and $\sum_{i=0}^{\infty} a_i = 1$.

Remark. Exponential smoothing applies a declining set of weights to all past data.

Remark. Multi-step-ahead and one-step-ahead forecasts are identical in case of exponential smoothing.

Exponential smoothing

One could fit the exponential curve $g(i) = \alpha \exp(-\alpha i)$ to the weight function $\alpha(1-\alpha)^i$. Hence, the "exponential smoothing" name



Forecast error for moving averages

Recall that for stationary time series,

$$D_t = \mu + \epsilon_t \tag{13}$$

hwere $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ For period t,

$$\begin{aligned} e_t &= F_t - D_t \\ &= \frac{1}{N} \sum_{i=t-N}^{t-1} D_i - D_t \end{aligned}$$

This leads to $\mathbb{E}\{e_t\} = \frac{1}{N} \sum_{i=t-N}^{t-1} \mathbb{E}\{D_i\} - \mathbb{E}\{D_t\} = \frac{N\mu}{N} - \mu = 0$.
MA(N) is unbiased.

Forecast error for moving averages

How about the variance?

$$
\begin{aligned}
Var(F_t - D_t) =& Var(F_t) + Var(D_t) \\
=& \frac{\sum_{i=t-N}^{t-1} Var(D_i)}{N^2} + Var(D_t) \\
=& \frac{N\sigma^2}{N^2} + \sigma^2 \\
=& \sigma^2 \frac{N+1}{N}
\end{aligned}
$$

Since $D_t \sim \mathcal{N}(\mu, \sigma^2)$, forecast error $e_t \sim \mathcal{N}(0, \sigma^2 \left(\frac{N+1}{N}\right))$

Forecast error for exponential smoothing

Recall

$$F_t = \sum_{i=0}^{\infty} \alpha(1-\alpha)^i D_{t-i-1}$$

This means,

$$\begin{aligned}
\mathbb{E}\{F_t - D_t\} =& \mu \sum_{i=0}^{\infty} \alpha(1-\alpha)^i - \mu = 0 \\
Var(F_t - D_t) =& \alpha^2 \sigma^2 \sum_{i=0}^{\infty} (1-\alpha)^{2i} + \sigma^2 \\
=& \frac{\alpha^2 \sigma^2}{1-(1-\alpha)^2} + \sigma^2 = \frac{2\sigma^2}{2-\alpha}
\end{aligned}$$

Remark.

1. ES(α) is unbiased.
2. Since $D_t \sim \mathcal{N}(\mu, \sigma^2)$, forecast error $e_t \sim \mathcal{N}(0, \frac{2\sigma^2}{2-\alpha})$

Moving averages versus exponential smoothing

- A moving average of order N and exponential smoothing with smoothing constant α would have the same distribution of error if

$$\frac{2}{2-\alpha} = \frac{N+1}{N} \implies \alpha = \frac{2}{N+1}$$

  Remark. This does not mean that forecasts obtained by both methods are the same.
- Both methods assume stationary process which may not be always true.
- Both methods will lag behind the trend if one exists.
- Both methods are based on weighted average of past data.
- ES is weighted average of all past data whereas MA is average of only last N periods of data.
- ES requires storing only the most recent data point (while moving averages requires storing the N most recent data points).

Outline

Trend-based forecasting methods

1. Regression analysis
2. Double exponential smoothing (Holt's method)

Linear regression analysis

1. State the problem
2. Model specification
   - An equation linking response and explanatory variables
   - Probability distribution of response variables
3. Parameter estimation (e.g., using Maximum Likelihood Estimation (MLE))
4. Check model adequacy (how well the model fits and summarizes the data)
5. Inference (for frequentist approach, we create confidence intervals, test hypothesis, and interpret results)

State the problem

We have data on response variable $\{Y_i\}_{i=1}^n$, which is to be predicted by independent variables $\{x_i\}_{i=1}^n$. Create a linear relationship between Y and x.

Model specification

Let us have the following relationship between response variables and predictor variables:

$$Y_i = \beta_0 + \beta_1 x_i + e_i$$

where, β_0 is the intercept and
β_1 is the coefficient of x, and e_i is the random error. Since, we do not expect every point to fall on the line, there would be some error equal to $e_i = Y_i - \beta_0 - \beta_1 x_i$. Let's assume this random error e_i follows i.i.d. normal distribution with mean 0 and variance σ^2, i.e.,

$$e_i \sim \mathcal{N}(0, \sigma^2)$$

Then,

$$Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$$

Parameter estimation

We need to estimate unknown parameters β_0 and β_1 in our model. One of the most common technique is maximum likelihood estimation (MLE). This is because ML estimates have good properties such as consistency, unbiassness, and efficiency. We can write the likelihood as below:

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(Y_i - \beta_0 - \beta_1 x_i)^2} \tag{14}$$

Taking the log of the above function:

$$l(\beta_0, \beta_1) = \log(L(\beta_0, \beta_1)) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i)^2 \tag{15}$$

To maximize this function, equate its gradient equal to zero.

$$\frac{\partial}{\partial\beta_0}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i)^2 = 0 \tag{16}$$

$$\frac{\partial}{\partial\beta_1}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i)^2 = 0 \tag{17}$$

Parameter estimation

This would give you a system of equations, which can be solved for β_0, β_1. Notice the least square term in (15). We are minimizing the sum of squares of error between given and predicted value.

$$\boxed{\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} x_i Y_i \end{bmatrix}} \tag{18}$$

We will not cover model adequacy and inference steps. For more interests, pick any book on statistical inference.

Double exponential smooting (Holt's method)

- It uses two smoothing parameters: α for the value of the series (intercept) and β for the trend (the slope).

$$S_t = \alpha D_t + (1-\alpha)(S_{t-1} + G_{t-1}) \tag{19}$$

$$G_t = \beta(S_t - S_{t-1}) + (1-\beta)G_{t-1} \tag{20}$$

  where, S_t is the value of intercept at time t and G_t is the value of the slope at time t.
- The first equation is the convex combination of most current observation of demand D_t and the prior forecast (previous intercept plus slope)
- The second equation can be explained as follows. Our new estimate of intercept S_t causes us to revise our estimate of slope G_t using convex combination of change in intercept and the prior slope.
- τ-step-ahead forecast made in period t is given by

$$F_{t,t+\tau} = S_t + \tau G_t \tag{21}$$

Outline

Season-based forecasting methods

A seasonal series is one that has a pattern that repeats every N periods. We refer to the number of periods (N) before the pattern begins to repeat as the length of the season.

Seasonality is specified by associating a multiplier with each period such that c_t, for $t = 1, \cdots, N$, such that $\sum_t c_t = N$. It indicates the average amount that the demand in period t is above or below the overall average. For example, $c_3 = 1.25$ means that the demand in the third period of the season is 25 % above the overall average demand.

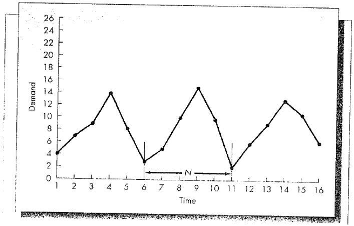


Figure: A seasonal demand series[3]

[3]Source: Chapter 2, Production and Operations Management by Steven Nahmias, 2018

Seasonal factors for stationary data

1. Compute the sample mean of all the data.
2. Divide each observation by the sample mean (this gives a seasonal factor for each period of observed data)
3. Average the factors for like periods within each season; these averages correspond to the N seasonal factors.

Winter's method

- Winter's method is type of triple exponential smoothing method.
- We assume a demand process of the following form

$$D_t = (\mu + Gt)c_t + \epsilon_t \tag{22}$$

  where μ is the base value (value of the intercept at $t = 0$), G is the slope (trend), $c_t$is the seasonality factor for the t-th period in the season,ϵ_t is random noise.
- We also assume that the length of the season is N, so that $\sum_t c_t = N$, and that c_t values are the same in each season.

Winter's method

- In each period t, S_t, G_t, and c_t are updated as follows:
  1. The series
  $$S_t = \alpha(D_{t-N}/c_{t-N}) + (1-\alpha)(S_{t-1} + G_{t-1}) \tag{23}$$
  2. The slope
  $$G_t = \beta(S_t - S_{t-1}) + (1-\beta)G_{t-1} \tag{24}$$
  3. The seasonal factors
  $$c_t = \gamma(D_{t-N}/S_t) + (1-\gamma)c_{t-N} \tag{25}$$
- The forecast made in period in period t for a future period $t+\tau$:
$$F_{t,t+\tau} = (S_t + \tau G_t)c_{t+\tau-N} \tag{26}$$

for $\tau \leq N$(for $t > N$, the seasonality factor would of the form $c_{t+\tau-iN}$ with $i = 2, 3, \cdots$)

Outline

Other methods we did not cover

- Box-Jenkins method (exploits possible dependencies (autocorrelation) among values of the series from period to period).
- Simulation for complex scenarios
- Machine learning models (neural network, etc.)
- Discrete choice (can be used when past data is not available)

Thank you!