Statistical Analysis

Section II : Basics of Time Series & Forecasting

COINS

Fall 2019

Forecasting Data and Methods

The appropriate forecasting methods depend largely on what data are available.

1. If there are no data available, or if the data available are not relevant to the forecasts, then **qualitative** forecasting methods must be used. These methods require the use of well-developed judgemental forecast methods.
2. **Quantitative** forecasting can be applied when two conditions are satisfied:

i. numerical information about the past is available;
ii. it is reasonable to assume that some aspects of the past patterns will continue into the future.

Forecasting Data and Methods

Quantitative forecasts

Most quantitative prediction problems use either time series data or cross-sectional data (collected at a single point in time).

In this module we are concerned with forecasting future data, and we concentrate on the time series domain.

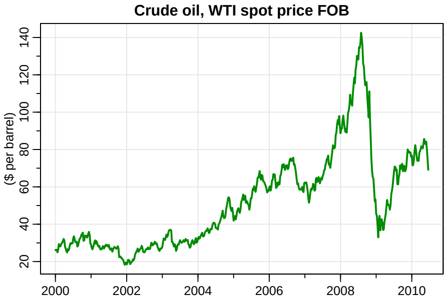
Time Series Explained

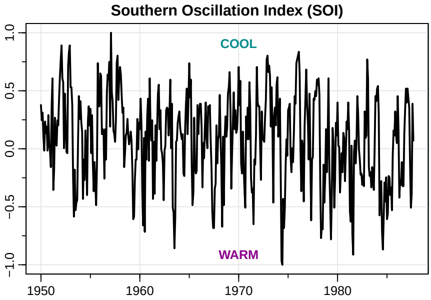
> *Often in forecasting, a key step is knowing when something can be forecast accurately, and when forecasts will be no better than tossing a coin. Good forecasts capture the genuine patterns and relationships which exist in the historical data, but do not replicate past events that will not occur again.* ~*Rob Hyndman*

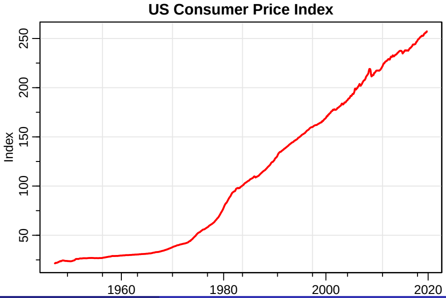
A time series is a sequence of measurements over time, usually obtained at regular, equally spaced intervals

- Every minute
- Hourly
- Daily
- Weekly
- Monthly
- Quarterly
- Yearly

Examples of Time Series Models







Components of a Time Series

- **Trends** (exists when there is a longrun increase or decrease in the data.)
  - Linear
  - Nonlinear
- **Seasonality** (occurs when a time series is affected by seasonal factors such as the time of the year or the day of the week.)
  - Seasonality is **always of a fixed and known frequency**.
    - These patterns repeat themselves within a year.
    - These fluctuations are usually due to economic conditions, and are often related to the "business cycle".
    - The duration of these fluctuations is usually at least 2 years.
- **Cycles**
  - Rises and falls that are not of a fixed frequency

Components of a Time Series

Additive

If we assume an additive decomposition, then we can write

$$y_t = S_t + T_t + R_t,$$

where y_t is the data, S_t is the seasonal component, T_t is the trend-cycle component, and R_t is the remainder component, all at period t.

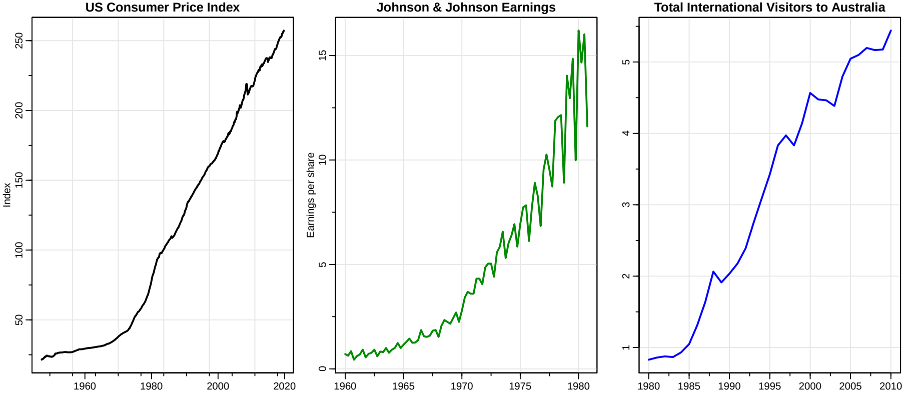
Multiplicative

A multiplicative decomposition would be written as

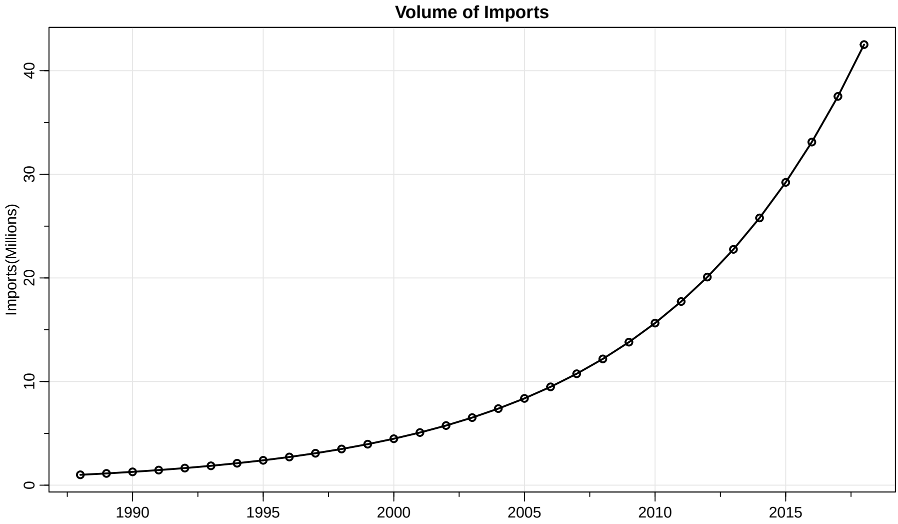
$$y_t = S_t \times T_t \times R_t,$$

Linear trend models

$$Y_t = a + b \cdot t + e_t$$



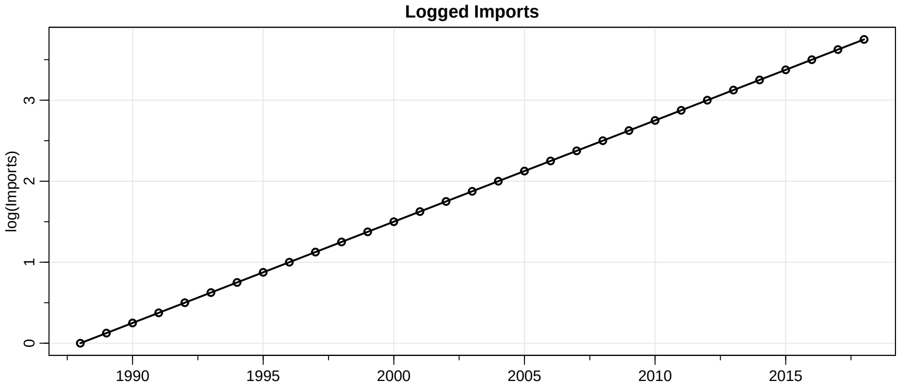
Nonlinear Trend



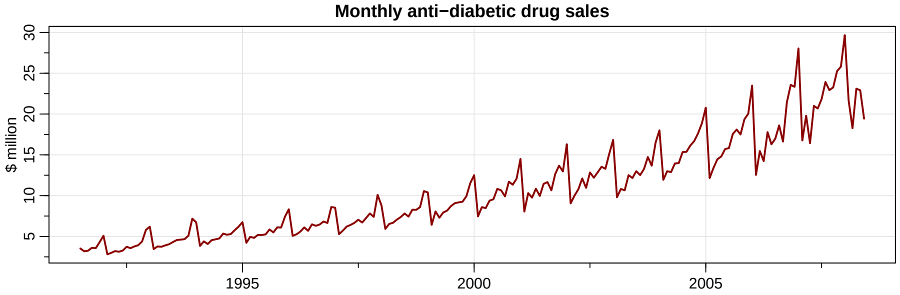
Transforming time series with nonlinear trends

$$\log Y_t = a + b \cdot t + e_t$$

Our transformed data now looks like:



Models with Trends and Seasonality



We can see a clear and increasing trend. There is also a strong seasonal pattern that increases in size as the level of the series increases.

The sudden drop at the start of each year is caused by a government subsidization scheme that makes it cost-effective for patients to stockpile drugs at the end of the calendar year.

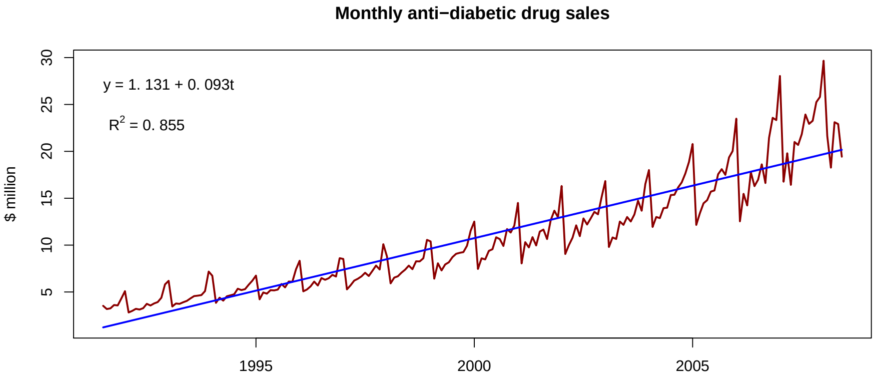
Any forecasts of this series would need to capture the seasonal pattern, and the fact that the trend is changing slowly.

Model Building

For the Monthly Anti-diabetic drug sales data:

- What kind of trends do you notice?
    - Linear
    - Logarithmic
    - Polynomial
- How can we smooth the data?
- How do we model the seasonality observed in the data?

Monthly Anti-diabetic drug sales data with Linear Trend



Immediately we see that a simple time trend explains almost 86% of the variations in the monthly sales of this anti-diabetic drug.

Modeling a nonlinear trend

Increasing Series

If the series appears to be changing at an **increasing rate** over time, we can use a model with a log transformation of the **data** as we learned earlier.

$$\log Y_t = a + b \cdot t + e_t$$

Equivalently, we can use an exponential model

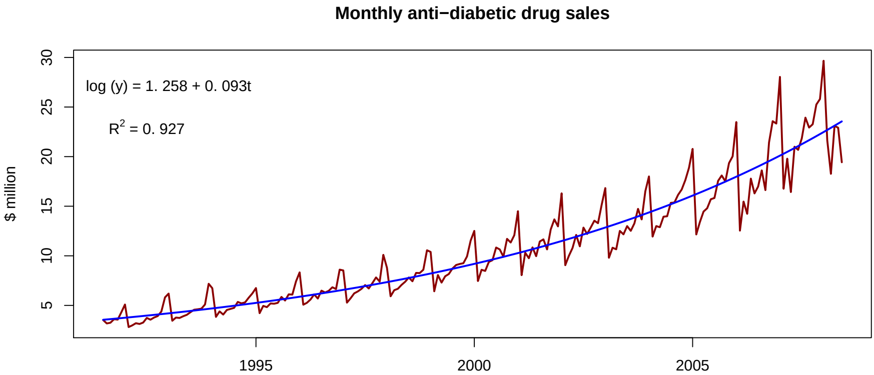
$$Y_t = \exp\{a + b \cdot t + e_t\}$$

Decreasing Series

If the series appears to be changing at an **decreasing rate** over time, we can use a model with log transformation of the **time trend**.

$$Y_t = a + b \cdot \log(t) + e_t$$

Monthly Anti-diabetic drug sales data with an Exponential Trend



From the R^2 we can conclude that the exponential trend does a good job of explaining the variations in the monthly sales of this anti-diabetic drug.

Modeling a Nonlinear Trend

- It is quite common to model trends using polynomials of varying orders:
- First Order (Linear)

$$\log Y_t = a + b \cdot t + e_t$$

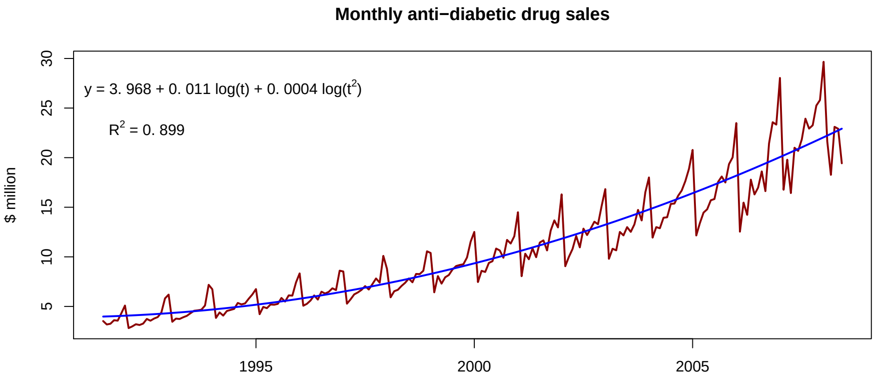
- Second Order (Quadratic)

$$\log Y_t = a + b \cdot t + c \cdot t^2 + e_t$$

- Third Order (Cubic)

$$\log Y_t = a + b \cdot t + c \cdot t^2 + d \cdot t^3 + e_t$$

Monthly Anti-diabetic drug sales data with a Quadratic Trend



Moving Average

- Moving averages are one of the core indicators in technical analysis, and there are a variety of different versions.
- One of the most common ways we examine trend in market data is using simple moving averages.
- In this approach, we forecast the value of the current time period using the last m consecutive observations.
- For example, a 4-point moving average would be computed as:

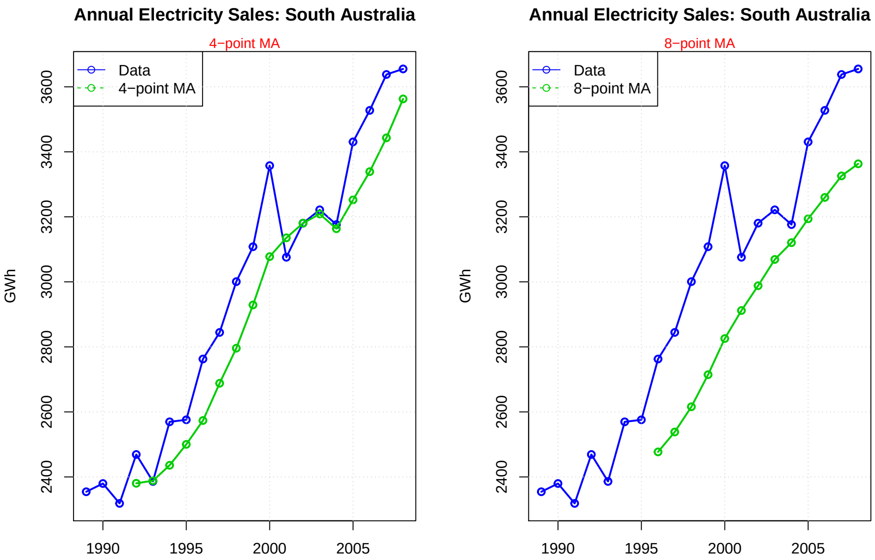
$$\bar{y}_{MA(4)} = \frac{y_t + y_{t-1} + y_{t-2} + y_{t-3}}{4}$$

- The average eliminates some of the randomness in the data, leaving a smooth trend-cycle component.

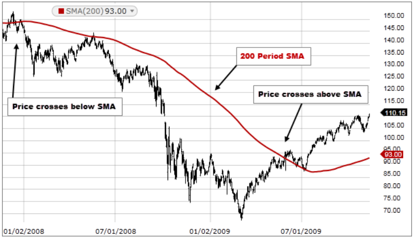
Moving Average

Table 1: Annual electricity sales to residential customers in South Australia. 1989–2008

Year	Sales (GWh)	4-point
1989	2354.34	
1990	2379.71	
1991	2318.52	
1992	2468.99	2380.39
1993	2386.09	2388.328
1994	2569.47	2435.768
1995	2575.72	2500.068
1996	2762.72	2573.5
1997	2844.50	2688.103
1998	3000.70	2795.91
1999	3108.10	2929.005
2000	3357.50	3077.7
2001	3075.70	3135.5
2002	3180.60	3180.475
2003	3221.60	3208.85
2004	3176.20	3163.525
2005	3430.60	3252.25
2006	3527.48	3338.97
2007	3637.89	3443.043
2008	3655.00	3562.743


Annual Electricity Sales: South Australia
4−point MA
Data
4−point MA
GWh
2400
2600
2800
3000
3200
3400
3600
1990
1995
2000
2005
Annual Electricity Sales: South Australia
8−point MA
Data
8−point MA
GWh
2400
2600
2800
3000
3200
3400
3600
1990
1995
2000
2005

Simple Moving Averages



MAs will prove very useful in your technical analyses of commodity prices. We often use them to:

1. Determine the direction of a trend. If the moving average is rising, it makes sense to buy, if it is falling, it's better to sell.
2. Round periods Moving Averages (50, 100, 200) can be taken as dynamic support and resistance levels.
3. The Moving Average Convergence Divergence (MACD) uses fast- and slow-moving averages to get a signal of a possible trend reversal, whether it's bullish or bearish.

Exponential Moving Averages (EMA)

- An exponential moving average is a weighted average that assigns positive weights to the current and past values of the time series.
- It gives greater weight to more recent values, and **the weights decrease exponentially** as the series goes farther back in time.
  - Data points closest to today's value will do a better job at explaining today's price so we will assign them greater weights than past data further away.

$$
\begin{aligned}
S_1 &= Y_1 \\
S_t &= wY_t + (1-w)S_{t-1} \\
&= wY_t + w(1-w)Y_{t-1} + w(1-w)^2 Y_{t-2} + \dots
\end{aligned}
$$

Exponential Moving Averages (EMA)

If $w = 0.5$

$$
\begin{aligned}
S_1 &= Y_1 \\
S_2 &= 0.5Y_2 + (1-0.5)S_1 = 0.5Y_2 + 0.5Y_1 \\
S_3 &= 0.5Y_3 + (1-0.5)S_2 = 0.5Y_3 + 0.5\big[(0.5)Y_2 + 0.5Y_1\big] \\
&= 0.5Y_3 + 0.25Y_2 + 0.25Y_1 \\
S_4 &= 0.5Y_4 + (1-0.5)S_3 = 0.5Y_4 + 0.5\big[0.5Y_3 + 0.25Y_2 + 0.25Y_1\big] \\
&= 0.5Y_4 + 0.25Y_3 + 0.125Y_2 + 0.125Y_1
\end{aligned}
$$

Exponential Moving Average (EMA)

Exponential Smoothing

- The weight we chose will affect the smoothness of the predicted value.
  - Smaller w's results in a smoother plot of predicted values.
  - As w gets closer to 1, the closer the prediction is to the original series.

Measuring Forecast Errors

- An actual value of time series observed at time t

$$Y_t$$

- A forecast value of Y_t

$$\widehat{Y}_t$$

- The forecast error

$$e_t = Y_t - \widehat{Y}_t$$

- The best model should minimize forecast errors

$$\min_{\hat{Y}_t} \sum_{t=1}^{T} e_t \tag{1}$$

Measuring Forecast Errors

$$\min_{\hat{Y}_t} \sum_{t=1}^{T} e_t \tag{1}$$

The potential problem with Eqn(1) is that forecasts with large errors but different signs can give minimum errors. For example, for two forecast errors, (-6,6), the sum of these errors is zero. For two forecast errors, (1,1.2), the summation of these errors are 2.4. The latter should be chosen as better forecasts. However, we will end up choosing the former one.

Next, we will look at some of the methods that we use to better inform our judgement.

Measuring Forecast Errors

- Method (1) - Mean Absolute Deviations (MAD)

$$MAD = \frac{1}{T}\sum_{t=1}^{T}\left|Y_t - \widehat{Y}_t\right|$$

- Method (2) - Mean Squared Error (MSE)

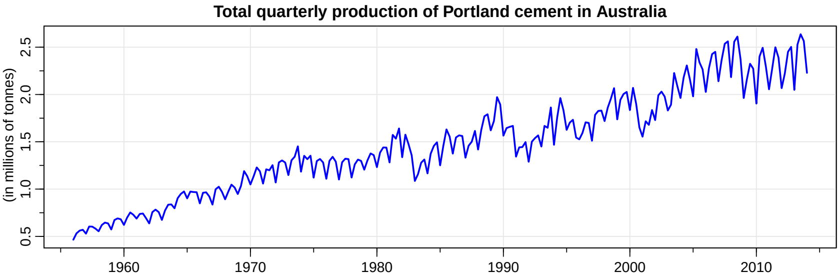
$$MSE = \frac{1}{T}\sum_{t=1}^{T}\left(Y_t - \widehat{Y}_t\right)^2$$

- Method (3) - Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(Y_t - \widehat{Y}_t\right)^2}$$

Modeling Seasonality

- The simplest way to model seasonality is using a dummy or indicator variable in a regression model.
- Let us attempt this for a quarterly data series of cement production.



Clearly, there is a trend and seasonality. As such, we can model this data as:[1]

$$Y_t = a + bt + c_1 Q_1 + c_2 Q_2 + c_3 Q_3 + e_t \tag{2}$$

[1] Notice we left out one of the dummies. This is important in order to avoid the *dummy variable trap*.

Modeling Seasonality

$$Y_t = a + bt + c_1 Q_1 + c_2 Q_2 + c_3 Q_3 + e_t \tag{2}$$

here

$$Q_1 = \begin{cases} 1 & \text{if quarter 1} \\ 0 & \text{if quarter 2,3,4} \end{cases}$$

$$Q_2 = \begin{cases} 1 & \text{if quarter 2} \\ 0 & \text{if quarter 1,3,4} \end{cases}$$

$$Q_3 = \begin{cases} 1 & \text{if quarter 3} \\ 0 & \text{if quarter 1,2,4} \end{cases}$$

Cement Production Data forecasted using Trends and Seasonality